Syllable-Based Speech Recognition

prepared for:

Speech Research Group
Personal Systems Laboratory
Texas Instruments, Inc.
PO Box 655474, MS 238
Dallas, Texas 75265

by:

J. Hamaker, A. Ganapathiraju and J. Picone
Institute for Signal and Information Processing
Department of Electrical and Computer Engineering
Mississippi State University
Box 9571
413 Simrall, Hardy Rd.
Mississippi State, Mississippi 39762
Tel: 601-325-3149
Fax: 601-325-3149
email: {hamaker,ganapath,picone}@isip.msstate.edu

EXECUTIVE SUMMARY

Large vocabulary continuous speech recognition (LVCSR) is the focal task for speech researchers in the world today. The most prominent approach to this problem is the use phones for modeling of spoken words. However, time and research have shown that phones are too small an acoustical unit to model temporal patterns and variations in continuous speech. Thus, a need exists for a new technique capable of exploiting both the spectral and temporal characteristics of continuous speech.

In response to this need, the Institute for Signal and Information Processing (ISIP) at Mississippi State University has developed an approach to LVCSR, using syllables as the fundamental acoustic unit, which is capable of overcoming these significant issues. The syllable's appeal lies in its close connection to articulation, its integration of some co-articulation phenomena, and the potential for a relatively compact representation of conversational speech. Our research is particularly timely as it seems phone-based approaches have achieved limited success in tasks such as SWITCHBOARD.

This research effort involves exploration of the syllable as a unit of recognition in the context of the SWITCHBOARD task initiated by ARPA and DOD. Our success on this task involved the following three phases: development of strong training and test data sets, design of phone and syllable baseline systems, exploration of new techniques to accentuate the strengths of syllable-based modeling. Of primary interest were integration of finite-duration modeling and monosyllabic word modeling. Results for this task are summarized below.

System	Word Error Rate
Context-Independent (CI) Monophone	62.3
Context-Dependent (CD) Phone	49.8
632 CI Syllables with 200 Monosyllabic Words, CD Phones and a Finite-Duration Topology	**49.1**

As an addition to the SWITCHBOARD task we have also developed a speaker-independent alphadigit recognition system applicable to many telephony applications. The alphadigit recognizer makes use of the phone-based system designed as part of the SWITCHBOARD system. As such, there is wide latitude remaining for tuning of this system, particularly performance on the E-set (B,C,D,E, etc.). However, performance (12.2% word error rate) is already on par with many state-of-the-art alphadigit recognizers (typically performing at 10-20%) using telephone data.

We have provided TI with all components of our SWITCHBOARD syllable experiments and Alphadigit systems. A TCL-TK graphical user interface (GUI) demo for demonstration of the system performance is also available. In addition, we have made all of the experiments, tools, data, etc. for both the syllable and alphadigit recognition systems available in the public domain (http://isip.msstate.edu/projects/lvcsr/).

TABLE OF CONTENTS

1. ABSTRACT

In the latter half of the twentieth century, the problem of large vocabulary continuous speech recognition (LVCSR) has been the focal point of speech research. Wide-ranging applicability and scope have kept LVCSR research at the fore-front of technology. To date, the most successful systems have used phone-based Hidden Markov Model (HMM) technology. Despite their successes, these systems still fall far short in performance. Triphones are a relatively inefficient decompositional unit due to the large number of frequently occurring patterns. Also, the triphone is not suitable for integration of spectral and temporal dependencies because it spans an extremely short time-interval.

A larger acoustical context is necessary for applications such as SWITCHBOARD. We believe the syllable can provide this context. Its primary appeal lies in its close connection to articulation, its integration of some co-articulation phenomena, and the potential for a relatively compact representation of conversational speech. This document presents a new set of experiments exploring the use of syllable-based acoustic models in LVCSR applications. In addition, we develop state-of-the-art systems for both LVCSR and Alphadigit recognition.

2. HISTORICAL BACKGROUND

The motivation for our work this summer can be traced as far back as the 1950s with the implementation of phonetic typewriters by Fry and Denes and Dreyfus and Graf. [1][2] Though these devices fell far short of their expectations, solving the “speech problem,” the framework we use today can be linked to their efforts. Their goal was to “use acoustic features of speech and a knowledge of phonetics to turn flowing speech into phonetic representations.” [3] In a broad sense, this can still be seen as the goal of speech recognition scientists throughout the world.

Based on this ground-breaking work, researchers at Bell Labs, RCA, and MIT used the 1960s as a testing ground for more expansive and complex systems. Systems using dynamic time warping, and pattern matching were introduced which performed well on small isolated word tasks. [4] However, they were still unable to make good progress into LVCSR systems.

In the 1960s, 70s and 80s, impressive strides were made in the field which pushed it in a new direction. Groups at T. I., I.B.M., and others began releasing commercial products in speech recognition. In the 60s and 70s these products were able to solve what is now the somewhat trivial recognition of discrete utterances in a relatively noise free environment. These were somewhat specialized and had limited vocabularies, but by the late 1980s systems existed which were capable of recognizing a vocabulary of 20,000 words spoken in isolation.

Speech research in these times was characterized by a movement from template-based technology to statistical modeling approaches. This is, in a large way, thanks to the advances in digital computing over the years. Scientists were able to develop more complex, more robust, and yet less expensive systems. Thus they were able to attempt recognition strategies which would have been impossible with the old analog and digital systems. Digital computers allowed us to replace the deterministic systems of the 1950s and 60s with stochastic processes based on probabilistic models.

In the 1990s speech recognition research has continued to move forward and to stay on the cutting edge of technology. The use of statistical modeling has become prevalent in the industry. Some scientists have also begun to explore concepts in neural networks. Realizing that the best speech recognition machine (the human) uses a network of this type, they are trying to mimic this ability using artificial neural networks. To date, this approach lags far behind the use of statistical modeling approaches, but continues to make advances.

In addition to advancements in computing and science, funding agencies have played large roles in speech research. In the late 1980s ARPA programs were established to give the industry a push toward continuous speech and large vocabulary applications. These efforts, and ARPA's, support have extended into the 1990s and is a driving force yet today. In recent years, ARPA has supported many large programs such as the Wall Street Journal, HUB-3 and HUB-4 and SWITCHBOARD corpora and evaluations. In these evaluations, the most prominent research groups around the world test the performance of their best systems on a task set by ARPA. Recent results on SWITCHBOARD as part of the Hub-5e evaluations are promising as shown in Table 1.

Laboratory	Word Error Rate
Carnegie Mellon - ISL	35.1%
BBN	35.5%
Cambridge - HTK	39.2%
Dragon	39.9%
BU	41.5%
SRI	42.5%

Table 1: Synopsis of recent ARPA evaluations.

Another major impetus behind advancements in the speech recognition community recently has been the summer workshops sponsored by the Department of Defense (DoD). Over the past five years, workshops have been held at Rutgers University in 1993 and 1994, and at Johns Hopkins in 1995, 1996, and again this year. The workshops are an invited research workshop on innovative techniques for LVCSR application. The invitees span industry, government and academe, including groups from Cambridge University, M.I.T., Entropic, Johns Hopkins, Oregon Graduate Institute, DoD, AT&T, SRI, Carnegie-Mellon, a host of others, and in 1997 our group from Mississippi State. Each of these groups brings with them insight into and experimental solutions for the problems facing LVCSR today. Recent workshops have dealt with corpora such as TIMIT, Wall Street Journal, and SWITCHBOARD and the exploration of new technologies including language modeling for conversational speech [5], use of discourse analysis to model pronunciation variations [6], RASTA processing of speech [7], and syllable-based speech processing [8]. It is with the latter of these that ISIP has become heavily involved.

3. SYLLABLE-BASED SPEECH RECOGNITION

For at least a decade now the triphone has been the dominant method of modeling speech acoustics. However, triphones are a relatively inefficient decompositional unit due to the large number of frequently occurring patterns. Moreover, since a triphone unit spans an extremely short time-interval, it is not suitable for integration of spectral and temporal dependencies. For applications such as SWITCHBOARD, where performance of phone-based approaches is unsatisfactory, the focus has shifted to a larger acoustic context. The syllable is one such acoustic unit. Its appeal lies in its close connection to articulation, its integration of some co-articulation phenomena, and the potential for a relatively compact representation of conversational speech.

We also conjecture that using a syllable as the fundamental acoustic unit obviates the need for explicit pronunciation modeling, since it can model many of the common variations in pronunciation based on a longer context window. Furthermore, an analysis of the hand-transcribed data from the SWB corpus [9] revealed that the deletion rate for syllables was below 1%. Not surprisingly, the comparable rate for phone deletions was an order of magnitude higher — 12%. This is a clear indication of the stability of a syllable-sized acoustic unit.

The use of an acoustic unit with a longer duration also makes it possible to simultaneously exploit temporal and spectral variations. Parameter trajectories [10] and multi-path HMMs [11] are examples of techniques that can exploit the longer acoustic context, but have had marginal impact on triphone-based systems. Recent research on stochastic segment modeling of phones [12] demonstrates that recognition performance can be improved by exploiting correlations in spectral and temporal structure. However, these experiments were limited to phone-based systems — their viability on larger units is yet to be proven. We believe that applying these ideas to a syllable-sized unit, which has a longer contextual window, will result in significant improvements.

4. THE SWITCHBOARD TASK

The SWITCHBOARD task is a part of an on-going set of government sponsored workshops and evaluations based on the T. I. SWITCHBOARD database. This task is driven by ARPA’s desire for more robust very large vocabulary telephone-based continuous speech recognition systems. ISIP is participating in this effort as part of the DOD-sponsored 1997 John’s Hopkins Summer Speech Workshop. Many of the best minds in speech technologies have come together at this workshop in a cooperative effort to better the speech recognition technology currently available.

Our goals in the SWITCHBOARD task have been five-fold this summer:

- Develop a robust and well-documented set of SWITCHBOARD-specific utilities. This includes data preparation, training, testing, etc. To date, this has been something which was lacking from the workshops. There were many copies of utilities designed by an assortment of groups, but with no commonality among them, and most were not well documented. We desired to build a suite which could be used and easily modified by any and all.

- Develop a baseline phone-based LVCSR system. This system would be used as a point of comparison for new technologies developed at the Workshop.

- Develop a baseline syllable-based system which demonstrates the syllable as a plausible fundamental unit for LVCSR applications.

- Develop a graphical demonstration of the systems we have designed.

- Delivery of this new technology to T. I. and the public domain so that they might investigate the use of the systems we designed for future applications.

Of course, we did not start these tasks from scratch. There was a fairly good groundwork set for us that had been laid out by previous workshops. Particularly, we were delivered the data and training and testing scripts which had been used at last years workshop. However the state of these scripts was not conducive to the robust research environment we desired. Thus, to help us reach our goals and to provide a world-class system to T. I. and the JHU workshop, ISIP put forth a plan to give a face-lift to both the data and supporting utilities. This plan is examined in further sections of this document.

4.1. Development of Data

The SWITCHBOARD corpus was developed in 1993 by T. I. through ARPA sponsorship as one of a series which included the TIMIT [13], Resource Management [14], and ATIS [15] corpora. SWITCHBOARD addresses the need for large multi-speaker telephone bandwidth speech. [16] For this reason, it has quickly become one of the most widely used LVCSR corpora in this time. Recent speech workshops and evaluations have been based solely around this data and it has become an invaluable source of information for researchers around the world.

SWITCHBOARD consists of 2500 conversations from approximately 500 speakers spanning both genders and every major dialect in American English. Each conversation is from 3 to 10 minutes so that a total of approximately 250 hours of data and 3 million words of text is available in this corpus. Each conversation is fully transcribed as a sequence of turns as shown in Figure 1. [17]

```
2627-A    96.4450    97.1150 [SILENCE]  OH YEAH [SILENCE] (2627-A-0027)

          2627-A:    <Conversation No.>-<Conversation Side>
         96.4450:    Beginning time of utterance in conversation
         97.1150:    Ending time of utterance in conversation
      “OH YEAH”:     Transcription of utterance
    2627-A-0027:     <Conversation No.>-<Conversation Side>-<Utterance No.>
```

Figure 1. Example transcription of SWITCHBOARD data. Note that the original transcriptions we received were marked with beginning and ending sample number rather than the time markings.

The corpus was automatically collected on T1 telephone lines. Each side of the conversation was recorded on a separate channel using 8 bit Mu Law at an 8 kHz sample rate. The audio quality of the data is generally good. There is a variable level of cross-talk in the data as well as the expected non-speech noises produced by the speakers. Table 2 gives a brief synopsis of the corpus.

During the 1996 JHU Workshop, the SWITCHBOARD data was segmented into a training set of speakers and a test set of speakers. These speaker sets did not overlap since a speaker-independent system was under development. We adopted these same speaker sets for our purposes as they gave us the coverage we desired and allowed us to be consistent with the ‘97 Workshop.

While we were satisfied with the makeup of the training and testing data, we determined that more work needed to be done on the data to bring it in line with our expectations. Namely, we desired a pristine training database, and a consistent set of transcription schemes. In addition, many groups released corrections to the original data which needed to be incorporated into the training transcriptions. Before development of our systems could begin, it was necessary to clean up the data, not only for our use, but for the use of others at the Workshop. The following items, detail the steps we took to replace the ‘96 Workshop data with an improved set.

- We built a new training set using the various correction sets (particularly the one developed at SRI) where applicable and the BBN transcriptions otherwise.

- We incorporated a consistent set of transcription schemes throughout the corrected set and the BBN transcriptions.

- We split the training data into clean and unclean sets, removing all utterances containing non-speech noises and word fragments from the clean data. This new clean data set became the official training set for both our work and for others at the Workshop. The unclean data was retained for historical purposes. Figure 2 gives examples of the unclean utterances removed and Table 3 gives the statistical makeup of our final data sets.

- We developed the official phone and syllable-based lexicons for the training and test data. This was an important step because training and evaluation would be based on the indications in these lexicons. Thus, we coordinated this closely with Dr. Barbara Wheatley, an expert in this area, with the Department of Defense. These lexicons can be found on our web-site.

```
2005-B    458.0350    458.5450  SURE [CROSS_TALK] (2005-B-0079)

2325-B     70.4190     71.5385  [MOUTH_NOISE] YEAH [CROSS_TALK] (2325-B-0018)

3081-A     19.5330     21.3450  ((SALAD)) IN A (3081-A-0008)

3530-B    226.2904    228.2250  [LAUGHTER] GET AWAY FROM IT ALL (3530-B-0069)

4314-A     33.5750     34.6188  WOW THAT’S PRETTY WILD- ~ (4314-A-0016)
```

Figure 2. Some examples of unclean utterances removed from our SWITCHBOARD training data.

		Male	Female	Total
No. of Speakers		302	241	543
No. of Conversations Sides		3126	2606	5732
Dialect	New England	12	11	23
	North Midland	47	30	77
	Northern	51	27	78
	Western	45	40	85
	South Midland	76	84	160
	Southern	35	22	57
	New York City	22	11	33
	Mixed	12	15	27
	Unknown	2	1	3
Age	20-29	89	59	148
	30-39	100	82	182
	40-49	68	44	112
	50-59	40	48	88
	60-69	5	8	13
Education	No High School	12	2	14
	Some High School	9	31	40
	College	161	148	309
	More than College	118	58	176
	Unknown	2	2	4

Table 2: SWITCHBOARD corpus speaker demographics. A more detailed and cross-referenced version is supplied by NIST with the SWITCHBOARD corpus.

		Training			Testing		
		Male	Female	Total	Male	Female	Total
No. of Speakers		205	179	384	21	12	33
No. of Conversations Sides		1409	1585	2994	25	13	38
Dialect (# of conv sides)	New England	17	77	94	4	1	5
	North Midland	257	220	477	3	0	3
	Northern	197	227	424	4	0	4
	Western	190	226	416	1	3	4
	South Midland	418	536	954	6	5	11
	Southern	140	130	270	7	2	9
	New York City	124	90	214	0	1	1
	Mixed	66	78	144	0	1	1
	Unknown	0	1	1	0	0	0
Age (# of conv sides)	20-29	334	219	553	5	3	8
	30-39	501	640	1141	8	5	13
	40-49	241	293	534	5	1	6
	50-59	267	388	655	5	4	9
	60-69	66	45	111	2	0	2
Education (# of conv sides)	No High School	18	0	18	0	1	1
	Some High School	10	165	175	1	3	4
	College	798	1022	1820	10	7	17
	More than College	563	384	947	14	1	15
	Unknown	20	14	34	0	1	1

Table 3: Demographics of data in our final SWITCHBOARD data sets

The test data derived at the ‘96 Workshop was based on acoustic segmentation. In this set, a speaker’s turn in the conversation was determined by acoustic pauses and turn boundaries. We became involved in a new effort this year to develop a test set based on linguistic segmentations. Instead of using acoustic cues for segmentation we desired to use linguistic cues based on the word-level transcriptions. Figure 3 shows an example of the difference between the two. These changes were in cooperation with other members of the Workshop and required a short amount of time on our part wherein we listened to each utterance and ensured that linguistic boundaries of the utterances lined up with the acoustic information in the data. In other words we modified the time marks of the transcriptions such that the linguistic segmentations determined from text files matched with the acoustic data available in the wave files, thus eliminating words which had been cutoff by mis-labeling the time markings. In addition, we split utterances where a large stretch of silence (more than 2 seconds) was present in the linguistically segmented data. Upon doing this, our test data expanded from 2119 utterances to over 3000 utterances. Initial tests run by Bill Byrne at Johns Hopkins showed that using the newly segmented data could bring an approximately 2% absolute drop in the error-rate, thus making it worth the amount of time necessary for refinement of the linguistically segmented data.

Acoustic segmentation:
2131-A 233.4400 234.6200 [SILENCE] YEAH THEY’RE GOOD [SILENCE] (2131-A-0059)

Linguistic Segmentation:
2131-A 233.4400 233.8400 YEAH (2131-A-0058)
2131-A 233.8400 234.6199 THEY’RE GOOD (2131-A-0059)

Figure 3. Example of move from acoustic segmentation to a linguistic segmentation. Note that the utterance has been split between “YEAH” and “THEY’RE GOOD” where one would expect a change in pace and tone by the speaker and in this case a short pause as well.

4.2. System Selection and Refinement

The system we selected for our work with the SWITCHBOARD data was defined at the 1996 JHU Workshop and was slated to be used at the ‘97 Workshop as well. It is a speaker-independent, Hidden Markov Model-based system. The basic structure of the full recognition system is shown in Figure 4. We used the Entropic HTK package whose training and recognition facilities are interfaced by a myriad of scripts.

The interface to training of the HTK system had also been defined at the ‘96 Workshop. The interface had been based on a single shell script which controlled the entire training run. Though it was for the most part technically sound, was an impediment to quick development of new systems. Thus we determined that it was necessary to rewrite the entire interface based on the old script. This required a considerable amount of time since we first had to decipher the intentions of the original script and then document and rewrite it in a suitable form. This task was, however, well worth the effort since our revised scripts have become the backbone of many of the 1997 Workshop systems.

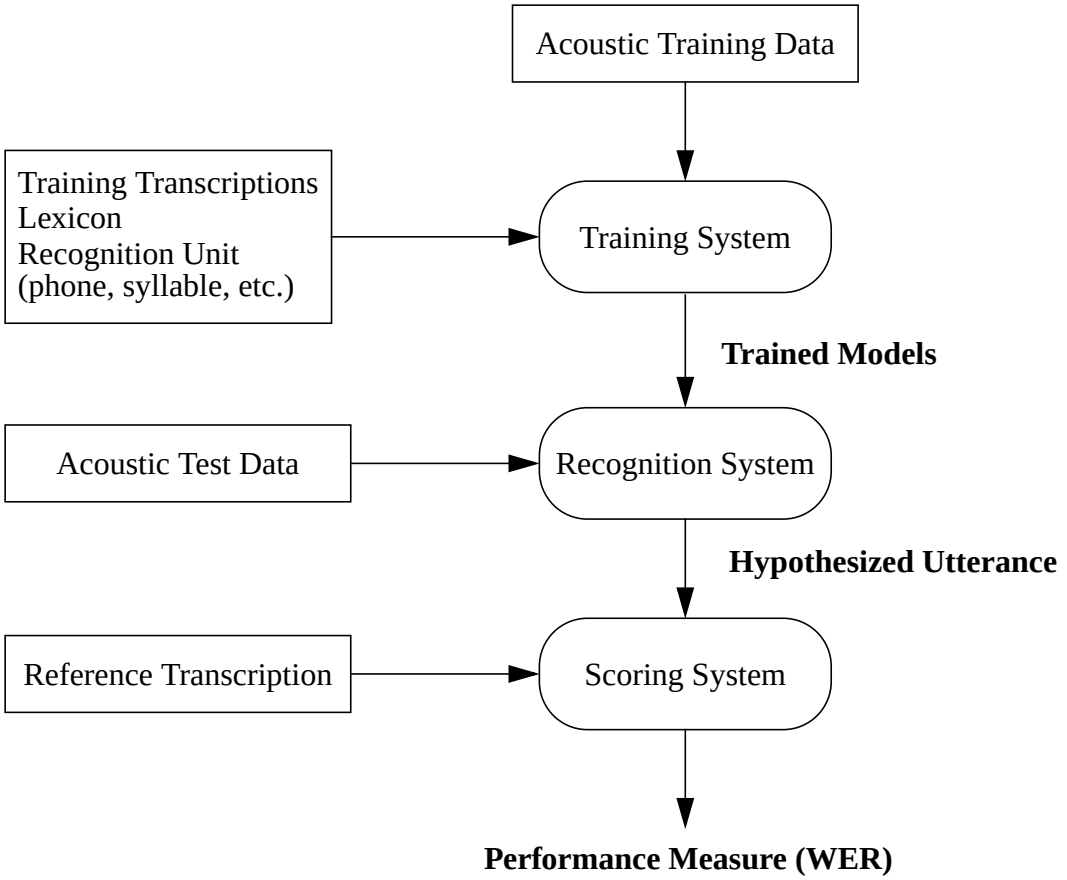

Figure 4. High-level view of the SWITCHBOARD recognition system.

Over the course of our investigation, we have found that research of the syllable as the fundamental unit of recognition was somewhat uncharted territory. Thus, we found it necessary to pursue many different concepts in our research. Some of these include finite duration modeling, hybrid (syllable and phone) systems and monosyllabic word modeling. Though our primary interest for the SWITCHBOARD and alphadigit tasks was to analyze the syllable it was also important to develop a baseline by which it could be compared. Thus, we also studied a basic phone system.

4.3. Phone-Based SWITCHBOARD System

Our phone-based system follows a fairly generic formula for triphone recognition. Figure 5 gives the detailed flow of this system. It is essentially a four-stage process consisting of:

- **Flat-start monophone training:** Generation of flat-start monophone seed models and reestimation of generated models.

- **Further training of monophones and forced alignment:** Correction of silence model, full training of single-Gaussian monophones, forced alignment of transcriptions to monophones, training with aligned transcriptions.

- **Initial triphone creation and training:** Creation of triphone transcriptions from monophone transcriptions, initial triphone training, triphone clustering, training of clustered triphones.

- **Creation and training of triphone mixtures:** Training of mixture.

Initially a context-independent monophone system was constructed. This system contains a phone set of 42 monophones, a silence model and a word-level silence model (short pause). All phone models were developed as 3-state left-to-right models without skip states. These models were seeded with a single Gaussian observation distribution. The number of Gaussians was increased to 32 per state during reestimation using a segmental K-means approach and a small amount of speech data.

Further training of the single-Gaussian monophone models was performed with a much larger training set consisting of 10 hours of data chosen to span the variations in the corpus. With this step came the addition of extra transitions to the silence model. These transitions provided a more robust model capable of absorbing the impulsive noises in the training data. In addition, a forced alignment of the monophone transcriptions was performed based on the fully-trained monophone models. The monophone models were retrained using these forced alignments.

A context-dependent phone system was then bootstrapped from the context-independent system. The single-Gaussian monophone models from the context-independent phone system were clustered and used to seed the triphone models. Four passes of Baum-Welch reestimation were used to generate single-component mixture distributions for the triphone models.

Finally, these models were increased to eight Gaussians per state using a standard divide-by-2 clustering algorithm. The resulting system had 81314 virtual triphones, 11344 real triphones, 34042 states and 8 Gaussians per mixture. The final count for the number of Gaussians is, however, reduced by tying states in the triphones.

Several features common in state-of-the-art SWITCHBOARD LVCSR systems were deliberately not included in this baseline system since the main goal of this work was to study the feasibility of syllables as an acoustic unit. In fact, it is hoped that some of these features will not be needed in a syllable system due to the inherent advantages of the syllable. The most prominent missing features were the use of a crossword decoder, a trigram language model, vocal tract length normalization, and speaker adaptation.

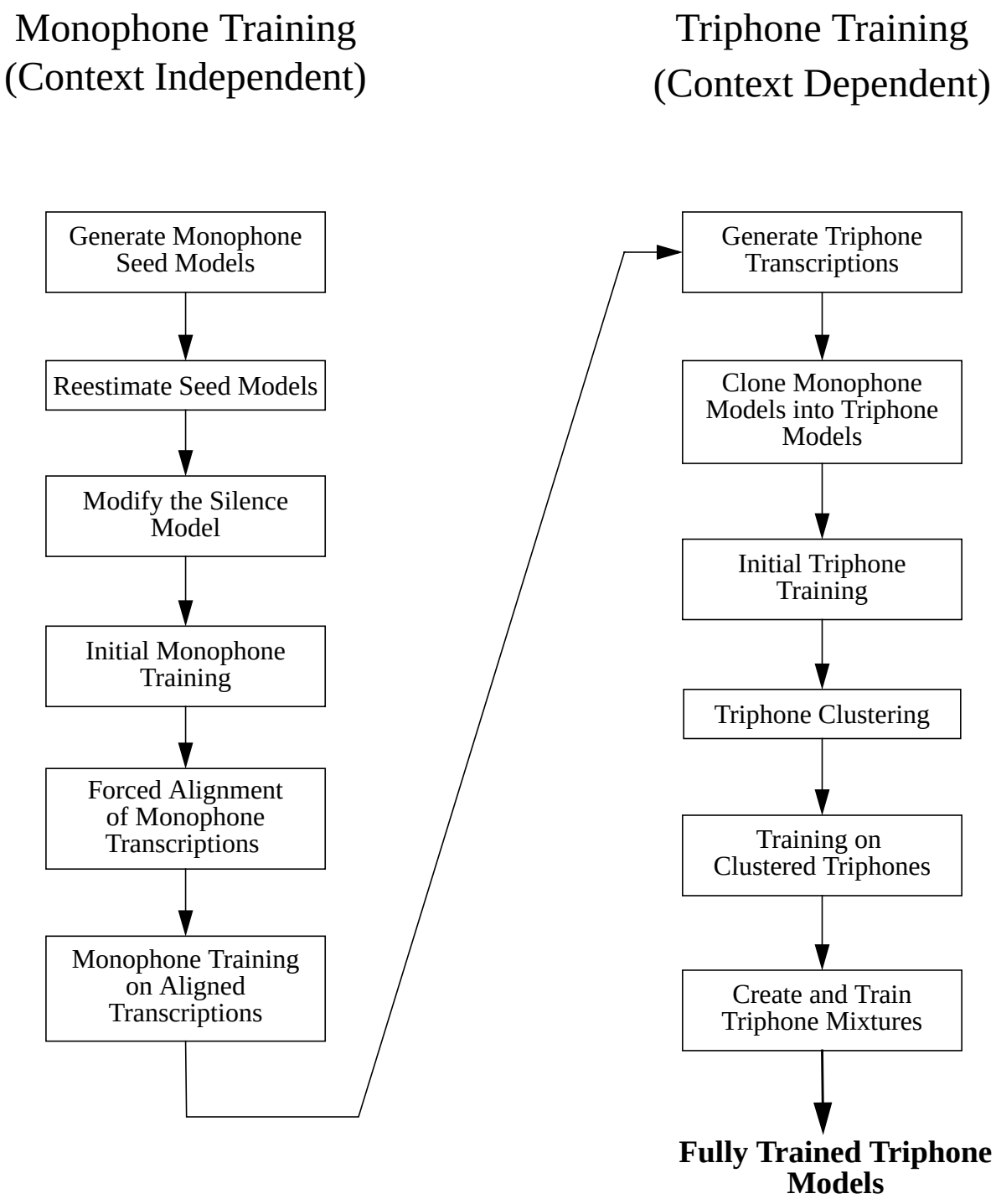

Figure 5. Detailed Flow of SWITCHBOARD Training System.

4.4. Syllable-Based SWITCHBOARD System

Perhaps the most critical issue in a syllable-based approach is the number of syllables required to cover the application. The number of lexical syllables in English is estimated to be on the order of 10000. This makes building a context-dependent syllable system a challenge. The first step in developing such a system was to represent each entry in the lexicon, previously defined in terms of phones, as a sequence of syllables — a process known as *syllabification* of the lexicon. We used a syllabified lexicon developed at Workshop'96 [18] for this stage. This lexicon consisted of over 70000 word entries for SWITCHBOARD and required 9023 syllables for complete coverage of the 60+ hour training data [9].

The model topology for the syllable models was kept similar to the context-dependent phone system. However, each syllable model was allowed to have a unique number of states. The number of states was selected to be equal to one half the average duration of the syllable, measured in 10 msec frames. The duration information for each syllable was measured from a forced alignment based on a state-of-the-art triphone system. Syllable models were trained in a manner analogous to the context-dependent phone system, minus the clustering stage. The resulting models had 8 Gaussians per state.

4.5. Hybrid Systems

Given the limited syllable coverage achievable in the baseline system, it was imperative that a system comprising a mixture of phones and syllables be developed to handle words not covered by the syllabary. To lower the computational expenses, this system was trained using a subset of the syllabary consisting of all syllables that occurred at least 20 times in the training database. This resulted in a set of 2419 syllables. We refer to this approach of mixing acoustic units as a hybrid system.

Since the hybrid system had both syllables and phones, each unique word in the training database could be classified into one of three categories — *syllable-only*: words which have one or more syllables in their lexical representation but do not have any phones, *phones-only*: words which have only phones in their lexical representation and *mixed*: words which are represented in terms of both phones and syllables.

It was observed that many models in the above system were poorly trained. This indicated a mismatch at the syllable phone junction. Due to time constraints, we circumvented this problem by building a system consisting of the 800 most frequent syllables and word-internal context dependent phones. It is interesting to note here that these 800 syllables covered almost 90% of the training data. The remaining 10% were replaced by its underlying phone representation. Several important issues, such as ambisyllabicity and resyllabification were ignored in this process. For example, if a syllable with an ambisyllabic marker was to be replaced by its phone representation, we ignored the marker all together. For instance,

sh ey d _#d_ih_ng $\Rightarrow$ sh ey d ih ng

The following example shows how the context for a sequence of phones was obtained from the adjoining syllables:

_ah_n k ⇒ _ah_n n-k

p _t_ih_ng ⇒ p+t _t_ih_ng

Syllable models from the above system and triphone models from the baseline triphone system were combined and reestimated using 4 passes of Baum-Welch over the entire training database.

4.6. Finite Duration Modeling

As mentioned before, we expect the syllable to be durationally more stable than the phone. However, when we looked at the forced alignments using our baseline system, we noticed very long tails in the duration histogram for many syllables. We also observed a very high deletion rate which was likely a result of the long durations. This suggested a need for some additional durational constraints on our models.

To explore the importance of durational models, we decided to evaluate a finite duration topology. A finite duration model was created by using the corresponding infinite duration model as a seed, and replicated each state in the finite duration model P times, where P is obtained using the following equation:

$$P = E[S] + 2.stddev(S) = f(p)$$

where S is the number of frames that have been mapped to that state for a given syllable token. Note that this computation is a function of p, the self-loop probability. The observations of each replicated state are tied to the observations of the entry state so that we maintain a manageable number of free variables for a model, and that there is sufficient training data per parameter. To achieve a quick turnaround time we decided not to do a complete training of the models. Rather, we did a 4 pass reestimation of the seed models from the baseline syllable and triphone systems.

4.7. Monosyllabic Word Modeling

In the systems described thus far, syllables were represented in a context-independent manner. This, however, may not be a good assumption for some or all the syllables. Syllables that exist as a monosyllabic word, and also appear as part of the pronunciation of another word demonstrate a much greater variation in pronunciation. We implemented a small number of monosyllabic word models as an attempt to capture some of this context dependency in syllables. Also, monosyllabic words constitute about 80% of the word tokens in the SWB corpus and accounted for 70% of the word error rate. The error rate on these words is about the same as the overall error rate. However, as a percentage of the total errors, monosyllabic words are clearly dominant. Hence, an experiment was conducted to create a separate model for each of the 200 most frequent monosyllabic words. We chose the 200 most frequent monosyllabic words and the remaining syllables which occurred more than 114 times in the training data. The remaining syllables were

expanded into word-internal triphones. The number, 114, is derived from the 800-syllable system whose least frequent syllable had 114 occurrences in the training data.

The training data for this system was aligned using the context-independent hybrid syllable system. The alignments were relabeled to reflect the 200 monosyllabic words. A new syllable inventory was defined in which a syllable was included based on the number of remaining training tokens after removing the monosyllabic words. The durations of syllables and words were then reestimated. The final system had 200 monosyllabic words, and 632 syllables and word-internal triphones.

5. EXTENSION OF SYSTEM TO ISOLATED ALPHADIGIT STRINGS

To obtain a measure of the extensibility of our continuous speech recognition systems developed on the SWITCHBOARD data, we have also developed models for a smaller scale Alphadigit vocabulary. The first reason that this was appealing was that the Alphadigit task was far removed from the task of continuous speech recognition. The corpus we were to choose would be comprised of short controlled segments of prompted speech, much different than the spontaneous unpredictable SWITCHBOARD utterances. This would give us an idea of the generality of our developed system. The second draw to this task was its real-world applicability. Telephone alphadigit recognition has been of interest to Bell Labs and others since the 1970's. [19] Many applications (security, automated telephone services, etc.) could be enhanced if a user's spelled or spoken response could reliably take the place of the keypads which are pervasive today. Thus, we felt that, with this task, we could deliver to T. I. a system with both practical and scientific application.

A robust and reliable alphadigit system has long been a goal for speech recognition scientists. Recent work on both alphabet and alphadigit systems has taken a focus on resolving the high rates of recognizer confusion for certain word sets. In particular, the E-set (B, C, D, E, G, P, T, V, Z, THREE) and A-set (A, J, K, H, EIGHT). The problems occur mainly because the acoustic differences between the letters of the sets are minimal. For instance, the letters B and D differ primarily in the first 10-20 ms during the consonant portion of the letter. [20] Many techniques have been used to counteract these similarities, such as inclusion of weighting functions in Dynamic Time Warping Analysis [21], and knowledge-based approaches [22]. The 1980s saw the use of HMM-based systems (first proposed by Rabiner and Wilpon [23]). Enhancements to the HMM-based systems have yielded the best performance to date. A summation of many important works in Alphadigit recognition is shown in Table 4.

Authors	Year	Bandwidth	Speaker Dependent	Speaker Independent
Rabiner, et al. [19]	1979	3.2 kHz	---	79.0
Rabiner and Wilpon [21]	1981	3.2 kHz	88.5	84.6
Rabiner and Wilpon [23]	1987	3.2 kHz	89.5	---
Euler et al. [24]	1990	3.2 kHz	93.0	---
Huang and Soong [25]	1990	3.2 kHz	90.0	---

Table 4: Historical overview of performance on the telephone Alphadigit task. Notice that the performance has not appreciably changed in almost twenty years.

5.1. Development of Data

One reason for the large amount of work on alphabet and alphadigit recognition is the wide availability of high quality corpora dealing with these topics. LDC and OGI, to name a few, have a large repository of corpora suitable for scientific research. The OGI Alphadigit Corpus is a recent release, bearing a large resemblance to the SWITCHBOARD Corpus. It, too, is a telephone database collected from volunteers. Training and testing on this type of data was to our advantage since many of the principle applications of alphadigit recognition would be in telephony. The approximately 3000 subjects of the Alphadigit Corpus were volunteers responding to a posting on the USEnet. The subjects were given a list of either 19 or 29 alphanumeric strings to speak. The strings in the lists were each six words long, and each list was "set up to balance phonetic context between all letter and digit pairs." [26] All totaled, there were 1102 separate prompting strings which gave a balanced coverage of vocabulary and contexts.

Use of the Alphadigit data with the HTK system was more cumbersome than expected. HTK is not able to directly interpret the data format in which the Alphadigit data is encoded (single-channel mu-law). Thus, a number of conversions were necessary to prepare the data in a format which HTK could understand. The necessary steps for preparation of the Alphadigits data are summarized in Appendix A and have also been made available to OGI and placed on our web-site. Once the conversion was completed, the data format (mfcc files) matched that of the SWITCHBOARD system.

We wished to have an even split in the training and test data amongst male, female, and adolescent speakers and also along geographic lines similar to the split we used with SWITCHBOARD. Since this information was not readily available, we spent a short time preparing the demographics for the database by listening to each speaker and classifying them accordingly. Of course, there was no way to determine the geographical representation of the database merely by listening to them, so we were forced to drop the notion of separation along geographic lines. We did however achieve a split of the data along gender and age lines. Thus, we classified each

utterance as a male, female, child, or unclassified utterance. As well as being summarized in Table 5 below, this information has been referred to OGI and made available on our web-site.

	Male	Female	Child	Unknown	Total
No. of Speakers	1419	1533	30	1	2983
No. of Utterances	35680	38585	795	29	75089
No. of clean utterances	25284	25700	477	29	51490

Table 5: Gender and age makeup of the OGI Alphadigits corpus. Notice that only approximately two-thirds of the data was found to be clean.

Determining the gender makeup of the database was important because we were able to use this information to determine the training and test data. As in the Switchboard experiments, we were developing a speaker-independent system. The impetus for determining the gender demographics of the data was so we could split it evenly across the training and test data - 50% of the males, females, and children in the training data, and 50% of each in the test data. This would ensure that we did not overtrain a particular voice-type, and, therefore, skew our recognition system toward one speaker type.

Another step in creating the training and test data was to eliminate bad utterances from the sets. We deemed “bad” to mean any data which was marked by OGI with breath noise, mouth noise, laughter, and other non-speech noises. In addition, we removed any utterances which contained words which were not in the intended lexicon. In other words, any utterance which was not solely comprised of alphadigits was removed from the training and test sets. Removal of these utterances left us with approximately two-thirds of the original data which was usable for training and testing. Separation of the training and test lists was a fairly simple process at this point. The remaining speakers were divided along gender and age lines and for every two speakers in a respective group, one was placed in the training set and one was placed in the test set. A summary of the gender/age distribution is shown in Table 6.

	Training				Testing			
	Male	Female	Children	Total	Male	Female	Children	Total
No. of Speakers	710	767	15	1492	709	766	15	1491
No. of Utterances	12866	12925	248	26038	12418	12775	229	25422

Table 6: Gender and age demographics of the Alphadigit training and test data.

5.2. System Selection and Refinement

Since we were using this task primarily as an extension to our Switchboard system, we attempted to keep the conditions between the two systems as similar as possible. The only major changes which were necessary were changes to the lexicons, and model lists. There was no need to maintain models or lexical entries for phonemes, triphones, syllables or words which would never occur in training, testing, or application, so we eliminated these from our training. We were able to eliminate 14 monophones, 55,000 triphones and a variable number of syllables (depending on the type of SWITCHBOARD syllable lexicon the alphadigit list is compared to) from our training models. This gave us moderate reductions in computation time, but more importantly removed superfluous information from our Alphadigit systems which had been necessary for the SWITCHBOARD system.

Again, to keep consistency, we attempted to alter the SWITCHBOARD training system as little as possible when creating the Alphadigits systems (see section 4.3 of this document for details on the SWITCHBOARD phoneme system). This emphasizes the need we had for recreating the training scripts in a more orderly fashion. Transfer of the SWITCHBOARD system to a new task with the old scripts would have been extremely time consuming and prone to error. With our restructured script system, we needed only to change a few pathnames in the scripts and edit the monophone and triphone lists as described earlier, thus creating the Alphadigits training system in a matter of minutes.

One option which was rejected was the use of word-models for alphadigit recognition. Word models would seem ideal since these systems are known to have a very high accuracy on small-vocabulary utterances. However, the drawback to word models is its lack of extensibility. If one were to add a word to the vocabulary or language model (in an automated telephony application, for instance) the recognizer would have to be completely retrained. Therein lies one advantage of subword units such as the syllable — they are easily migrated to a different vocabulary.

6. GRAPHICAL DEMONSTRATION

The most important aspect of a speech recognition system is not its performance on evaluation data, but rather its applicability to a real-world situation. To demonstrate this with respect to our systems, we have designed a graphical user interface (GUI) which provides a live-input interface to the array of systems we have developed. The user simply records an utterance using a microphone or input device of choice, and pushes a button. The GUI will then display the input waveform and the utterance which was hypothesized by the recognition system. In addition, future versions of the GUI will provide automatic alignment of the utterance and recognition hypothesis based on an endpointing program. A snapshot of the GUI window is shown in Figure 6.

We have developed the demo using TCL/TK and a parameter-file-driven approach. Each recognizer configuration can be defined by a set of parameters so that the user can change the way that the recognizer behaves on the fly. The demo is coded such that it is ignorant of the particular

type of recognizer being used. This makes it easy to add new recognition systems as they become available. One need only define an interface to the core of the demo.

In the upper lefthand corner of the GUI are controls for sound input and playback via **Record** and **Playback** buttons. On the upper right are the interface control buttons. The **Run** button enacts the recognition system. There is also a **Configure** button which allows one to change various options including which recognition parameter file to use. This allows the user to change the type of recognition being used from within the GUI with a minimal low-level knowledge of the system being used. The controls on the demonstration are, for the most part, self-explanatory but there is some on-line help suitable for a first-time user via the **Info** button.

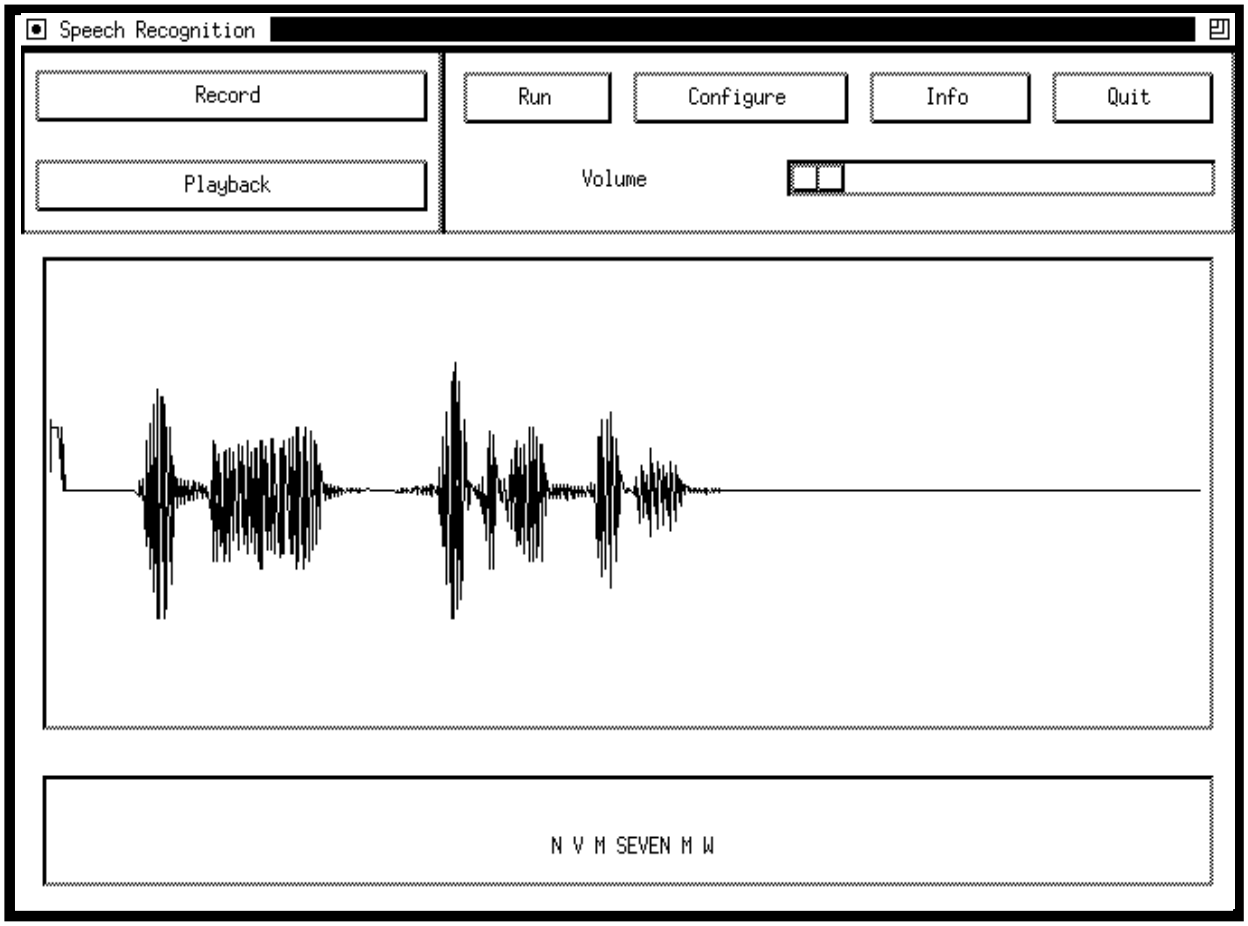

Figure 6. The graphical user interface for the Speech Recognition demonstration. Facilities are provided for live-input, playback of recording, file-driven recognition, and viewing of the waveform and recognizer hypothesis.

The bottom portion of the window is for the output display. There are two types of output available with this demo: the input waveform, and the recognition hypothesis. The image of the input waveform is generated assuming the data is in raw data format. Thus the user's input system should be capable of generating raw data files. The recognition hypothesis is read from the output of the recognizer through a global output file using the NIST standard "trn" file format. This ensures that the hypothesis display is not tied to any particular recognizer's output format.

7. SWITCHBOARD EVALUATION

Evaluation of our SWITCHBOARD systems used standard methods developed at last year's workshop and NIST evaluations. Scores were obtained through lattice rescoring. This rescoring was accomplished with the NIST standard utility *sclite* and the HTK decoders. The relevant statistic for recognition performance is the Word Error Rate (WER) and is determined by summing the percent of errors attributable to substitutions, insertions, and deletions. An example of the output from sclite is shown in Figure 7.

```
id: (2121-b-0032)
Scores: (#C #S #D #I) 5 1 0 1
REF:  uh overstated THAT a lot of *
HYP: uh overstated GOT    a lot of T
Eval:                  S                 I

id: (2131-b-0018)
Scores: (#C #S #D #I) 2 0 0 0
REF:   oh really
HYP:  oh really
Eval:

id: (2151-b-0017)
Scores: (#C #S #D #I) 2 0 1 0
REF:  BECAUSE that's just
HYP: *******     that's just
Eval:  D
```

Figure 7. Example of output alignments from *sclite*

7.1. Syllable Evaluations

Analysis of a baseline system was important as it gave us both an idea of our initial performance relative to phone systems and it provided insight into possible improvements. In addition, the process of creating a baseline system provided us with the opportunity to overcome some obvious obstacles. For instance, we initially used over 9000 syllables for full coverage of the corpus. This proved to be impractical due to computational constraints and undertraining of many syllables which occurred too few times in the data. To correct for this problem, we initiated a number of different systems which used a combination of high frequency syllables and phones for recognition.

In the first full evaluation, we attempted to develop and test fairly generic baseline phone and syllable systems. The syllable system consisted of the 800 most frequently occurring syllables and 42 monophones. Since our syllable system was a context-independent system, we also limited our phone systems to be context-independent. Results from these three experiments on the full test set of acoustically segmented data are shown in Table 7. The SYL0 system performs exactly as we would expect—between the context-independent monophone and context-independent triphone systems. This is an encouraging result as it shows the immediate utility of larger acoustical units such as the syllable.

System	Word-Error Rate (2192 Utterances)
Context-Independent Monophone	62.3%
Context-Dependent Phone	49.8%
Context-Independent Syllable and 44 Context-Independent Monophones (SYL0)	57.8%

Table 7: Results of baseline SWITCHBOARD experiments. Note that the baseline syllable system outperforms the comparable monophone-based system and falls short of the baseline triphone system.

Superficial analysis of the baseline system yields an important and intuitive observation: context-dependent phones are much better suited for a task such as SWITCHBOARD than are context-independent phones. With this in mind, we designed a second syllable system which incorporated these context dependent phones. This system known as SYL1 exhibited a substantial improvement over SYL0 of almost 6%. Results for this and the remaining syllable systems are shown in Table 9.

The SYL2 system is an extension of the SYL1 system with the inclusion of a finite duration topology as discussed in section 4.6. This provided another 2% decrease in word-error rate, but not in the same manner we had projected. Table 8 shows the insertion, deletion, and substitution rates for both the SYL1 system and the SYL2 system. Notice that, though we believed finite duration modeling would give improvement by decreasing the deletion rate, the gain was actually a result of decreasing the substitution rate. An explanation for this is not readily apparent but is a topic for future work. Another point of interest for the finite duration syllable system is the dramatic increase in system resources necessary to model this topology. We found that SYL2 would use on the order of 500 Megabytes of RAM which is as much as 4 or 5 times the memory used by SYL1. This is due to the increase of states necessary to model the finite durations.

System	WER	Substitutions	Insertions	Deletions
SYL1	51.7	**33.9**	3.5	**14.3**
SYL2	49.9	**32.3**	3.5	**14.1**

Table 8: Results of SYL1 and SYL2 experiments. Decrease in deletions is not the primary cause in improvement of WER as was anticipated.

The SYL4 system consists of 632 syllable models, 200 monosyllabic word models and context. As mentioned earlier, a large number of errors were from the confusability between the monosyllabic word and the corresponding syllable model. For example, the syllable model “_ae_n_d” or context-dependent phone model “ae-n+d” could be confused for the sound made by

the word “and” whose proper representation is the monosyllabic word model “__and”. We have not confirmed this trend, but we believe this is the reason for the lack of substantial improvement in SYL4. The small increase in performance we do see is likely due to the inherent ability of word models to do pronunciation modeling. Though the monosyllabic word models give only a marginal improvement, efficient modeling of these could have a profound effect on recognition.

Our final, best system was the SYL6 system which merged the techniques giving improvement in individual tests. Specifically, this was a duplicate of the SYL4 system cast into a finite duration topology. There is only a small increase in performance between SYL4 and SYL6, but the real comparison point is between SYL6 and the baseline context dependent phone system. The syllable system not only meets, but exceeds the performance of the comparable phone system.

System	Word-Error Rate
CI Monophones	62.3%
CD Monophones	**49.8%**
800 CI Syllables and CI 44 Monophones (SYL0)	**57.8%**
800 CI Syllables and CD Phones (SYL1)	**51.7%**
800 CI Syllables and CD Phones with Finite Duration Topology (SYL2)	49.9%
632 CI Syllables with 200 Monosyllabic Words and CD Phones(SYL4)	49.3%
SYL4 Models Duration split to long and short (SYL5)	49.5%
SYL4 Models converted to finite duration (SYL6)	**49.1%**

Table 9: Results of SWITCHBOARD Syllable experiments. The SYL6 performs better than the baseline triphone system. Also, note that the use of context-dependent phones, modeling of monosyllabic words, and finite duration modeling each yielded improvements in system performance. NOTE: CI and CD represent context-independent and context-dependent respectively

8. ALPHADIGITS EVALUATION

There was quite a bit of difference between the SWITCHBOARD and Alphadigit testing mechanisms. Two things, in particular, required change for the Alphadigit system: generation of the grammar network and compensation for confusion pairs. Recall that for the Switchboard system, a separate lattice was developed for each test utterance. For the Alphadigit task we took a different approach. Since the possibilities for unique utterances were fairly limited, we were able to define a simple global grammar for the task as shown in Figure 8. This grammar requires silence at the beginning and end of the utterance and allows an unlimited number of spoken alphadigits in between the silences.

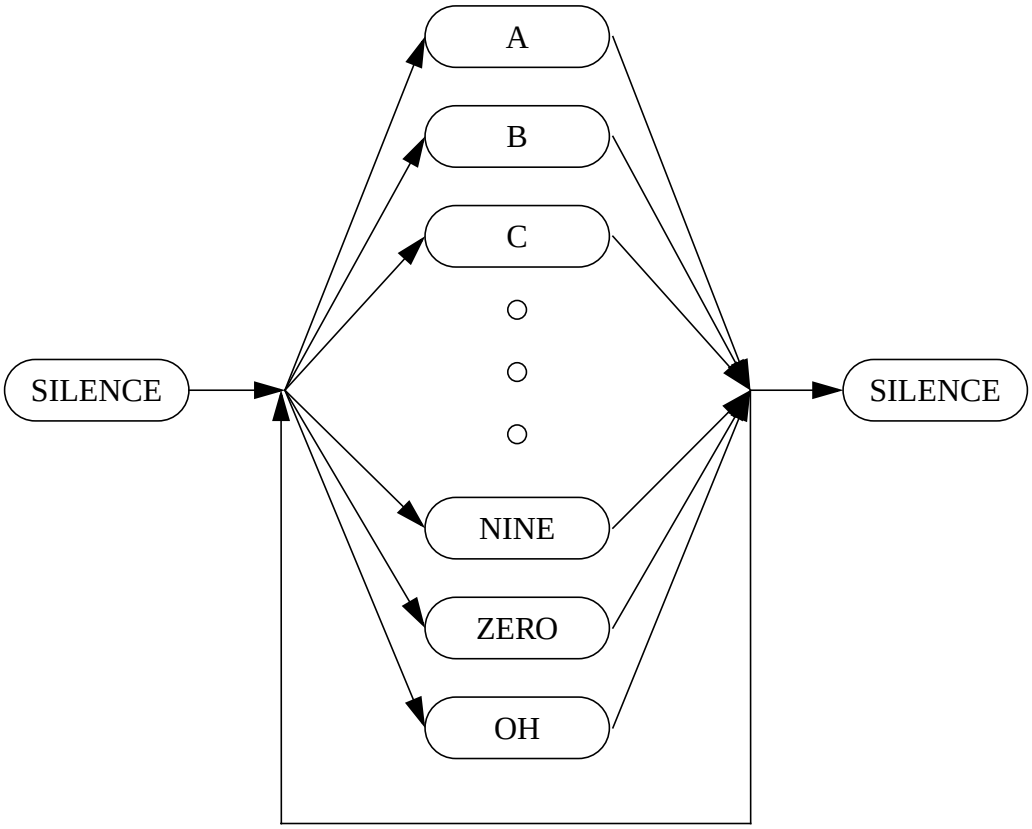

Figure 8. Diagram of the Alphadigit grammar network. The form of any utterance is: beginning silence, one or more alphadigits, followed by ending silence.

Also in testing was the need to account for the confusion pair of ‘o’ and ‘oh’. It was likely that the recognizer would substitute one for the other since they had the exact same phonetic makeup (represented by the phone ‘ow’). To avoid being penalized for this mistake we post-processed the recognition hypothesis and pre-processed the recognition reference transcriptions so that each instance of ‘oh’ was changed to ‘o’. The ‘oh’ to ‘o’ transformation was arbitrary and could have just as easily been ‘o’ to ‘oh’ since there were no complex syntactical rules for utterances.

8.1. System Evaluations

Evaluation of the Alphadigit system was meant to prove the extensibility of our SWITCHBOARD systems to a small vocabulary task. In doing so, we again began with baseline phone-based systems and planned to develop syllable systems from this. However, the performance of our baseline phone systems was surprisingly good as indicated in Table 10. Thus, rather than pursuing the syllable approach for alphadigits, we continued exploration of our phone-based systems.

Two primary systems were evaluated: a word-internal context dependent phone system, and a cross-word context dependent phone system. In essence, a word-internal system does not allow

triphones to span across words whereas a cross-word system does. Naturally, one would expect that the cross-word system would be a better performer as it allows for an extended context. An example of the different phone representations for the string “five m five” is shown in Figure 9.

String:	FIVE M FIVE
Word-Internal Expansion:	sil f+ay f-ay+v ay-v eh+m eh-m f+ay f-ay+v ay-v sil
Cross-Word Expansion:	sil-f+ay f-ay+v ay-v+eh v-eh+m eh-m+f m-f+ay f-ay+v ay-v+sil

Figure 9. Comparison of word-internal and cross-word representations and context. The cross-word system of training provides a much higher degree of context.

As shown in Table 10, there is a strange trends in the results of Alphadigit testing. Note that the word-internal system does not scale to training. In other words, the WER does not decrease as the training set increases (by 900%). In contrast, there is an improvement in the cross-word system of about 2% as training increases. The probable cause of this phenomena is the undertraining of the many triphone models in the crossword system for the short training set. As expected, the cross-word systems perform better than the best word-internal system. In fact, our best cross-word system compares very well to most state-of-the-art systems (including those in Table 4).

System	Word-Error Rate
Word-Internal Phones (Small Training Set) - WI0	15.8%
Cross-Word Phones (Small Training Set) - CW0	14.5%
Word-Internal Phones (Large Training Set) - WI1	15.8%
Cross-Word Phones (Large Training Set) - CW1	**12.2%**

Table 10: Results of Alphadigit experiments. Note that the performance of the best cross-word system is by far the best among our experiments.

As suspected, the E and A-sets played the largest role in the accumulation of errors. Table 11 gives an analysis of the confusion pairs present in the CW1 system. Notice that the E and A-set account for over 50% of the errors. The S-F confusion pair also had an unexpectedly high impact on performance caused by the similarity in the “short E” sound which dominates each.

Confusion set	# of occurrences (out of 1912 substitutions)	Percentage of total # of pairs
E-set B,C,D,E,G,P,T,V,Z,THREE	828	43.3%
A-set A,H,J,K,EIGHT	254	13.3%
S-F	160	8.4%
U-set Q,U,TWO	103	5.4%
M-N	88	4.6%
O-L	71	3.7%
I-set I,Y,FIVE,NINE	61	3.2%

Table 11: Confusion pairs in CW1. Confusion among the E and A-sets is dominant.

9. CONCLUSIONS

Large vocabulary continuous speech recognition (LVCSR) has application in nearly all aspects of our lives. While many feel that the future of LVCSR research is in triphone recognition, we feel that a movement toward syllable-based modeling will have more profound effects on the performance of modern recognition systems. Syllables are highly desirable since they allow one to model both temporal and spectral variations in the speech.

While there is much work remaining on the topic of syllable-based speech recognition, our systems serve as a strong starting point. We have presented results showing that the syllable can model continuous speech with accuracy as well or better than comparable phone-based systems. This, coupled with the ability to model temporal phenomena makes the syllable an interesting and promising topic for future work.

In addition to our work with syllables in LVCSR, we have (almost by accident) produced a state-of-the-art speaker-independent alphadigit recognition system. This could have wide-spreading application in telephony and automated services of all sorts. However, it can be improved greatly by taking into account and adapting for confusion in the E-sets and A-sets.

In conjunction with our efforts to produce high-quality public-domain research, all experiments, scripts, tools, etc. pertaining to this project are available in the public domain at *http://isip.msstate.edu/projects/lvcsr/*.

10. FUTURE DIRECTIONS

Though our work in syllables is promising, we have just scraped the surface of their potential. The system presented here is clearly deficient in a number of areas, including the representation of ambisyllabics in the lexicon, and the integration of syllable and phone models in a mixed word entry. It is important to note that, to this point, we have developed a very basic system. Our current system does not exploit the temporal modeling advantages inherent to the syllable. Progress on these topics is slow, as we are developing many of the techniques for the first time. We do believe, however, that the current system provides the proper framework to simultaneously exploit the temporal and spectral characteristics of the syllable by clustering or trajectory modeling. Preliminary results in this direction are promising.

Another important area for future research is the introduction of context-dependent syllables in a constrained way to keep the number of free variables in the system manageable. Note that the syllable systems presented here do not use any form of state-tying across models or states, yet contain fewer parameters than their comparable context-dependent phone systems. Hence, we believe that additional syllable models can be introduced without a significant increase in the overall system complexity.

11. ACKNOWLEDGEMENTS

First and foremost, we wish to express our gratitude to our sponsor, Texas Instruments, and in particular Dr. Jack Godfrey for their continued support of our work. We also wish to thank Dr. George Doddington, Dr. Barb Wheatley and Mark Ordowski for their guidance in development of the Switchboard training and testing strategies. We are also grateful to the participants of the Johns Hopkins 1997 Speech Recognition Workshop, especially Dr. Bill Byrne and Dr. Fred Jelinek, for their support and assistance in our efforts. Finally, we would like to acknowledge Dr. Ron Cole, Mike Noel, Don Colton, and Paul Hosom of the Oregon Graduate Institute for their assistance and collaboration on the OGI Alphadigits Corpus experiments.

12. REFERENCES

[1] Fry, D.B., and P. Denes. “The solution of some fundamental problems in mechanical speech recognition,” Language and Speech, vol 1, pp. 33-58, 1958.

[2] Dreyfus,-Graf, J. “Phonetograph Und Schwallellen-Quantelung,” Proceedings of the Stockholm Speech Communication Seminar, Stockholm, Sweden, Sept. 1962.

[3] Deller, John R., Jr., John G. Proakis, and John H. L. Hansen. *Discrete-Time Processing of Speech Signals*. New York: Macmillan Publishing Company, 1993.

[4] Rabiner, Lawrence R., and Bing-Hwang Juang. *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[5] http://www.clsp.jhu.edu/lm95/lm95_general.html

[6] http://www.clsp.jhu.edu/ws97/discourselm

[7] Hermansky, Hynek, and Nelson Morgan, "RASTA Processing of Speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, Oct. 1994, pp. 578-589.

[8] Ganapathiraju, A. et. al., "Syllable - A Promising Recognition Unit for LVCSR," Submitted to ASRU, 1997.

[9] Greenberg, S., "The Switchboard Transcription Project", *1996 LVCSR Summer Research Workshop*, Research Notes 24, CLSP, Johns Hopkins University, April 1997.

[10] Gish, H. and Ng, K., "Parameter Trajectory Models for Speech Recognition," *Proc. of the IEEE ICASSP '97*, Munich, Germany, April 1997.

[11] Korkmazskiy, F., et. al., "Generalized Mixture of HMMs for Continuous Speech Recognition," *Proc. of the IEEE ICASSP* '97, Munich, Germany, April 1997.

[12] Ostendorf, M., and Roukos, S., "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *IEEE Trans. on Acoust., Speech and Signal Proc., vol. 37, no. 12, pp. 1857-1869, December 1989.*

[13] Fisher, W.M., G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," *Proceedings of the DARPA Speech Recognition Workshop*, 1986.

[14] Price, P.J., W.M. Fisher, J. Bernstein, D.S. Pallett, "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," *Proc. of the IEEE ICASSP* '98.

[15] Hemphill, Charles T., John J. Godgrey, and George R. Doddington, "The ATIS Spoken Language Systems Pilot Corpus," *Proceedings of the DARPA Speech and Natural Language Workshop*, Pittsburgh, PA, June, 1990.

[16] Godfrey, John J., Edward C. Holliman, and Jane McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," in *Proceedings of the IEEE ICASSP '92.*

[17] Young, Stephen J., Philip C. Woodland, and William J. Byrne, "Spontaneous Speech Recognition for the Credit Card Corpus Using the HTK Toolkit," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, October 1994, pp. 615-621.

[18] Ostendorf, M., et. al., "Modeling Systematic Variations in Pronunciations via a Language-Dependent Hidden Speaking Mode," *1996 LVCSR Summer Research Workshop*, Research Notes 24, CLSP, Johns Hopkins University, April 1997.

[19] Rabiner, L. S. Levinson, A. Rosenberg, and J. Wilpon, “Speaker independent recoggition of isolated words using clustering techniques,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 336-349, Aug. 1979.

[20] Loizou, Philipos C. and Andreas S. Spanias, “High-Performance Alphabet Recognition,” *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 6, pp 430-445, Nov. 1996.

[21] Rabiner, L. and J. Wilpon, “Isolated word recognition using a two-pass pattern recognition approach,” in *Proceedings of the IEEE ICASSP* ‘81.

[22] Cole, R., R. Stern, and M. Lasry, “Performing fine phonetic distinctions: Templates vs. features,” in *Invariance and Variability of Speech Processes*, J. Perkell and D. Klatt, Eds. New York: Lawrence Erlbaum, 1986, pp. 325-341.

[23] Rabiner, L. and J. Wilpon, “Some performance benchmarks for isolated word speech recognition systems,” *Comput. Speech Language*, vol. 2, pp. 343-357.

[24] Euler, S., B. Juang, C. Lee, and F. Soong, “Statistical segmentation and word modeling techniques in isolated word recognition,” in *Proceedings of the IEEE ICASSP* ‘90.

[25] Huang, E. and F. Soong, “A probabilistic acoustic MAP based discriminative HMM training,” in *Proceedings of the IEEE ICASSP* ‘90.

[26] http://www.cse.ogi.edu/CSLU/corpora/alphadigit/

[27] Woodland, P., et. al., *HTK Version 1.5: User, Reference and Programmer Manuals*, Cambridge University Engineering Department & Entropic Research Laboratories Inc., 1995.

APPENDIX A. CONFIGURING ALPHADIGIT CORPUS FOR USE WITH HTK

Described here is the methods by which one may use the OGI Alphadigits Corpus with the HTK system. Conversion of the corpus is necessary since HTK is unable to handle single-channel mu-law (it can handle interleaved mu-law), which is the format of the corpus, and there are settings in the wave-file header which causes HTK to crash.

A.1. Correction of Wave-file Headers

We have found that the wav file NIST headers cause HTK and Entropic Waves to crash. Neither seem to understand what the file format of the wave file is. We have determined that this is partly (if not fully) caused by settings in the wav file NIST headers. Shown below are settings in the header from a wav file as originally distributed with the Alphadigits database (these are the settings we are interested in):

```
sample_byte_format -s6 mu-law
sample_n_bytes -i 2
sample_sig_bits -i 16
```

Now we show the settings as they appear in the “fixed” wave files:

```
sample_coding -s4 ulaw
sample_n_bytes -i 1
sample_sig_bits -i 8
```

One can see that we had to remove the sample_byte_format setting and replace it with the sample_coding setting. Also we had to change the sample_n_bytes from 2 to 1 and sample_sig_bits from 16 to 8 (since ulaw is single-byte data). Changing these was done with the following sequence of commands using standard NIST tools.

```
# remove the sample_byte_format tag
#
h_delete -F “sample_byte_format” “$current_ogi_file”;

# change the sample_n_bytes field from 2 to 1
#
h_edit -I sample_n_bytes=1 “$current_ogi_file”;

# change the sample_sig_bits tag from 16 to 8
#
h_edit -I sample_sig_bits=8 “$current_ogi_file”;

# add the sample_coding field and set it to ulaw
#
h_edit -S sample_coding=”ulaw” “$current_ogi_file”
```

A.2. Configuring HCopy

HCopy is the program supplied by HTK to convert these wave-files to mfcc format. However HTK only understands a limited number of file formats. HTK does support ulaw coding but only in interleaved two-channel data. Thus the Alphadigit single-channel ulaw data had to be converted to format which HTK could comprehend. We chose the 16-bit pcm format since it was supported by both HTK and NIST conversion tools. To conserve memory, we also chose to do the conversion on the fly. The files on disk did not change, the file was just converted in memory before being converted to mfcc format. This was done by setting the HTK parameter HWAVEFILTER in the HCopy configuration file. This configuration file is shown below.

```
SOURCEKIND      =  WAVEFORM
SOURCEFORMAT    =  NIST
ZMEANSOURCE     =  TRUE
TARGETKIND      =  MFCC_E_Z
TARGETFORMAT    =  HTK
TARGETRATE      =  100000
SAVECOMPRESSED  =  TRUE
WINDOWSIZE      =  250000.0
USEHAMMING      =  TRUE
PREEMCOEF       =  0.97
NUMCHANS        =  24
CEPLIFTER       =  22
NUMCEPS         =  12
ENORMALISE      =  TRUE
ESCALE          =  1.0
SAVEWITHCRC     =  TRUE
HWAVEFILTER     =  w_decode -o pcm $ -'
```

A.3. Running HCopy

HCopy was executed with the following command line:

```
HCopy -T 1 -C ogi_convert_config.text -S mfcc.list
```

Switch	Value	Comment
-T	1	Trace Level
-C	ogi_convert_config.text	Configuration Parameter File (shown in section A.2)
-S	mfcc.list	list of source wave files and destination mfcc files

The format for the mfcc.list file is shown below:

/d02/ogi/wav/1/AD-11/AD-11.p1.wav /d02/ogi/mfcc/1/AD-11/AD-11.p1.mfcc
/d02/ogi/wav/1/AD-11/AD-11.p2.wav /d02/ogi/mfcc/1/AD-11/AD-11.p2.mfcc
/d02/ogi/wav/1/AD-11/AD-11.p3.wav /d02/ogi/mfcc/1/AD-11/AD-11.p3.mfcc
/d02/ogi/wav/1/AD-11/AD-11.p4.wav /d02/ogi/mfcc/1/AD-11/AD-11.p4.mfcc

APPENDIX B. SOFTWARE SYNOPSIS

name: eval.sh

synopsis: eval.sh file_format ref_file_list filename1.hyp filename2.trn

description: This script converts the HTK format files into "trn" format recognized by the NIST's "sclite" scoring software and evaluates the recognition process. Output files containing evaluation data will have names which are extensions of the hypothesis file (eg. filename1.score.sys). A "trn" format equivalent of the input hypothesis file is also created.

options/arguments:

file_format: input hypothesis file formats (mlf_word, mlf_word_align, mlf_model_align, mlf_state_align, trans, trn)

ref_file_list: reference file list.

filename1.hyp: Recognition output from a recognizer

filename2.score: Reference file in NIST's "trn" format with cut ids and time segmentations.

name: score_miss_fa.pl

synopsis: score_miss_fa.pl list_file sclite_alignment_file

description: This script computes the miss and false alarm percentages of words in a specified list by processing an sclite alignment file.

options/arguments:

list_file: list of words to check

sclite_align_file: file containing sclite alignments

name: gen_mlf

synopsis: gen_mlf [options] file_in file_out

description: This program converts a transcription file in many formats into an mlf file. The mlf file is the HTK standard format file.

options/arguments:

-mode:	conversion mode (default = word.)
-lexicon:	lexicon to use to convert words (default = none)
-trans:	transcription file (default = none)
-filter:	a file containing the filter instructions (default = none)
-help:	display a help message
file_in:	input lists of files containing utterance ids
file_out:	output mlf file