1. ABSTRACT

This report describes the research effort of the "Syllable-Based Speech Processing" team participating in the 1997 Summer Workshop on Innovative Techniques for LVCSR. We present an attempt to model syllable-level acoustic information as a viable alternative to the conventional phone-level acoustic unit for large vocabulary conversational speech recognition (LVCSR). The motivation for this work was the inherent limitations of phone-based approaches, primarily the excessive number of commonly occurring patterns and lack of a mechanism for modeling larger scale temporal dependencies. In this report we present preliminary but encouraging results on Switchboard. Our syllable-based recognition system, developed in less than six months of work, exceeded the performance of a comparable triphone system both in terms of word error rate (WER) and complexity. The WER of the best syllable system reported here is 49.1% on a standard SWITCHBOARD (SWB) evaluation. Durational modeling and spectral clustering have also been explored in this context. Further, the advantages in explicitly modeling monosyllabic words was demonstrated. One of the highlights of this research is the numerous strategies developed to train and test systems with mixtures of acoustic units.

2. INTRODUCTION

For at least a decade now the triphone has been the dominant method of modeling speech acoustics for speech recognition. 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, such a unit is not suitable for integration of spectral and temporal dependencies. For applications such as SWITCHBOARD (SWB) [1] where performance of phone-based approaches is unsatisfactory [2], 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. Also, an analysis of the hand-transcribed data from the SWB corpus [3] 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 [4] and multi-path HMMs are examples of techniques that can exploit the longer acoustic context (but as yet have had marginal impact on triphone-based systems). Recent research on stochastic segment modeling of phones [5] 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 [7].

3. BASELINE SYSTEMS

As part of our preliminary work on syllable-based LVCSR, two baseline systems were constructed based on research conducted at prior CLSP workshops: a *context-independent monophone system* and a *word-internal triphone system*. Both these systems were carefully designed to provide state-of-the-art performance on a standard SWB task within the constraints of the technology used for implementation. All systems described in this report were based on a standard LVCSR training and test procedure for a commercially available package — HTK[10]. The baseline systems were also used to validate the training process and the training scripts which were used during the workshop. We decided not to incorporate cross-word context for the syllable system, since this adds significant complexity to the decoder and may mask the fundamental advantages of syllable-based speech modeling. We also restricted our experiments to a bigram language model which could be efficiently processed in a lattice re-scoring framework. Our recognition experiments were based upon re-scoring lattices generated from a more sophisticated recognition system prior to the workshop. These lattices, supplied to all participants at the workshop, had a word error rate (WER) of approximately 10%. Though not described in any detail here, we also built a cross-word triphone system as a reference point which delivered a 45.6% WER.

3.1. Phone-Based Baseline Systems

Since the syllable models in all systems described here were context-independent, a comparable context-independent phone , or monophone system was constructed as a baseline. This system used a phone inventory consisting of 42 phones and a silence model (in addition, a word-level silence model was used as well). All phone models were standard 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 re-estimation using a segmental K-MEANS approach.

To construct the context-dependent phone system, single-Gaussian monophone models generated from the context-independent system were clustered and used to seed triphone models. Four passes of Baum-Welch re-estimation were used to generate single-component mixture distributions for the triphone models. These models were then successively refined to have eight Gaussians per state using a standard divide-by-2 clustering algorithm. The resulting system had 81,314 virtual triphones, 11,344 real triphones, 34,042 states and 8 Gaussians per mixture. The final count for the number of Gaussians is, however, reduced by tying states in the triphones. This word-internal triphone system resulted in 49.8% WER.

Several features common in state-of-the-art SWB 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. The most prominent missing features were the use of **a crossword decoder, a trigram language model, vocal tract length normalization, and speaker adaptation**. 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.

3.2. Syllable-Based Baseline System

Perhaps the most critical issue in a syllable-based approach is the number of syllables required to give good coverage of the application. The number of lexical syllables in English is estimated to be on the order of 10,000 [9]. 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 (WS'96) for this stage [8]. This lexicon consisted of over 70,000 word entries for SWB and required 9,023 syllables for complete coverage of the 60+ hour training data.

The WS'96 lexicon indicated syllabification, but still represented all pronunciations in terms of phone sequences, for example:

BEFORE → [b . ax] [f ‘ ow r]

AFTER → [‘ ae [f] t . er]

The brackets indicate syllabification, and the punctuation marks indicate stress levels. The overlapping brackets indicate "ambisyllabic" consonants, i.e., consonants that can be considered to be members of both the preceding and the following syllable (or alternatively but equivalently, the syllable boundary can be considered to fall within the ambisyllabic consonant). Ambisyllabic consonants may occur frequently in rapid or casual speech, and the WS'96 lexicon was designed to represent the most casual syllabification; therefore, ambisyllabic consonants were rampant, occurring in 54% of the entries.

In converting these entries to representations in terms of syllables, one issue was the proper treatment of the stress information, i.e., whether to create separate syllable models for pronunciations differing only in stress, or whether to merge such pronunciations into a single model. For the initial baseline system, stress was ignored, partly simply to avoid additional complication, but also because the value of lexical stress information seemed questionable. For cases where stress has a strong effect on vowel quality, the stress marker is redundant; phoneme identity implicitly encodes stress, as in the "*ax*" in "*BEFORE*". More importantly, actual stress in continuous spontaneous speech is quite different from lexical stress; many "stressed" syllables are actually unstressed in fluent speech.

The second major issue in converting the lexicon was the treatment of ambisyllabic consonants, i.e., whether to assign such consonants entirely to one of the adjacent syllables or to continue to treat them as belonging to both syllables. In this case, the decision was to continue to treat them as belonging to both syllables. This avoided the difficulty of determining the best criterion for assigning the consonant to a single syllable (the preceding one? the following one? the one where it belongs in very deliberate speech, where there are no ambisyllabics?). More importantly, it maintained the distinction between syllables whose initial (final) consonant is shared with the adjacent syllable and those whose initial (final) is wholly within the syllable. To maintain this distinction in the lexicon, a special symbol (#) was introduced to mark ambisyllabics.

In the syllabified lexicon, then, the examples shown above were represented as:

BEFORE → _b_ax _f_ow_r

AFTER → _ae_f# _#f_t_er

(The initial underscore distinguished syllables containing only one phone from monophones, e.g., *_ay* vs. *ay*) Both the treatment of stress and the treatment of ambisyllabicity deserve further research, but alternatives were not explored during the workshop for lack of time.

The model topology for the syllable models was kept similar to the context-independent phone system (left-to-right models without skip states). 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 median duration of the syllable, measured in 10 ms frames. The duration information for each syllable was measured from a forced alignment based on a state-of-the-art triphone system developed during WS'96. 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. This system however suffered from a large number of very poorly trained syllables due to insufficient training data. To circumvent this problem we tested a system based on the 800 most frequent syllables. The syllables which occurred less frequently were replaced by their corresponding phonetic representation. The monophones were trained as an independent system with 32 Gaussian components per state. The performance of this system was 57.8% WER.

Given this promising result, and the unwieldy nature of using the full set of over 9,000 syllables, we decided to use the smaller set described above for all further experiments. One advantage of this approach is that each syllable model is guaranteed a reasonable amount of training data. Another reason for going to a smaller subset of syllables was that the 9023 syllable system was consuming over 400M of memory during the training phase that generates the 8 mixture models. This slowed down the training process considerably due to excessive swapping of data from secondary memory. One drawback of the reduced system is that it becomes imperative that a

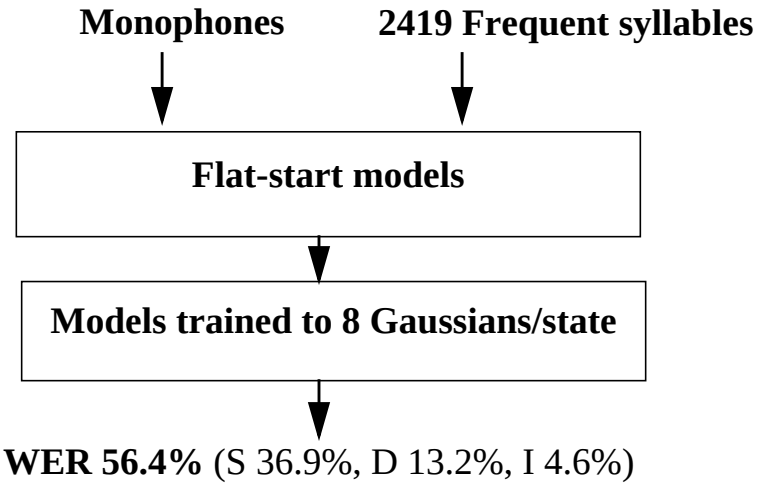

Figure 1: Baseline syllable system, 2419 syllables + 42 monophones

system comprised of a mixture of phones and syllables be developed to handle words not covered completely by the syllable inventory.

3.3. Hybrid System with 2,419 Syllables

As described in the previous section, the baseline syllable system was tested as a 800 syllable + monophone system in which syllables and monophones were trained independently. The next logical step was to train a system with the syllable and monophone parameters estimated concurrently. 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 training syllables and phones or any mixture of acoustic units (e.g. words and syllables) as a **hybrid system**.

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*" was represented as "*sh ey d_#d_ih_ng*." The problem with this treatment is that it replaces part of a phone, *d#*, which represents only the initial portion of the ambisyllabic *d*, with a model representing the entire phone *d*. In effect, then, the hybrid model contains one and a half consonants where only one should occur. The 2419 syllables and 42 phones in this system were trained together using training procedure described in Figure 1. This system delivered a 56.4% WER.

3.4. Hybrid System with 800 Syllables

It was observed that many models in the above system were still poorly trained. Due to time constraints, we circumvented this problem by building a system consisting of the 800 most frequent syllables and the 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. This system gave a performance of 55.1% WER.

Since the hybrid system had both syllables and phones, each unique word in the training database

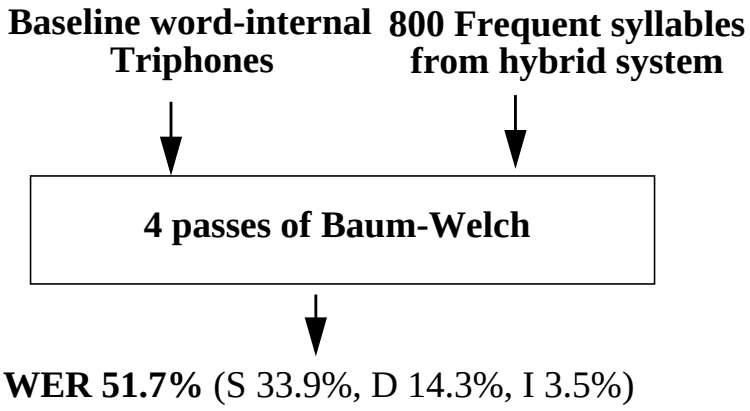

Figure 2: Hybrid syllable system with 800 syllables and word-internal triphones reestimated

Data set	# words	% miss	
		Baseline Syllable + Phone System	**Word-internal Triphone System**
All Words	18069	53	47
SO	15676	51	46
MX	1186	58	46
PO	1207	71	60

Table 1: Error analysis of the baseline syllable system

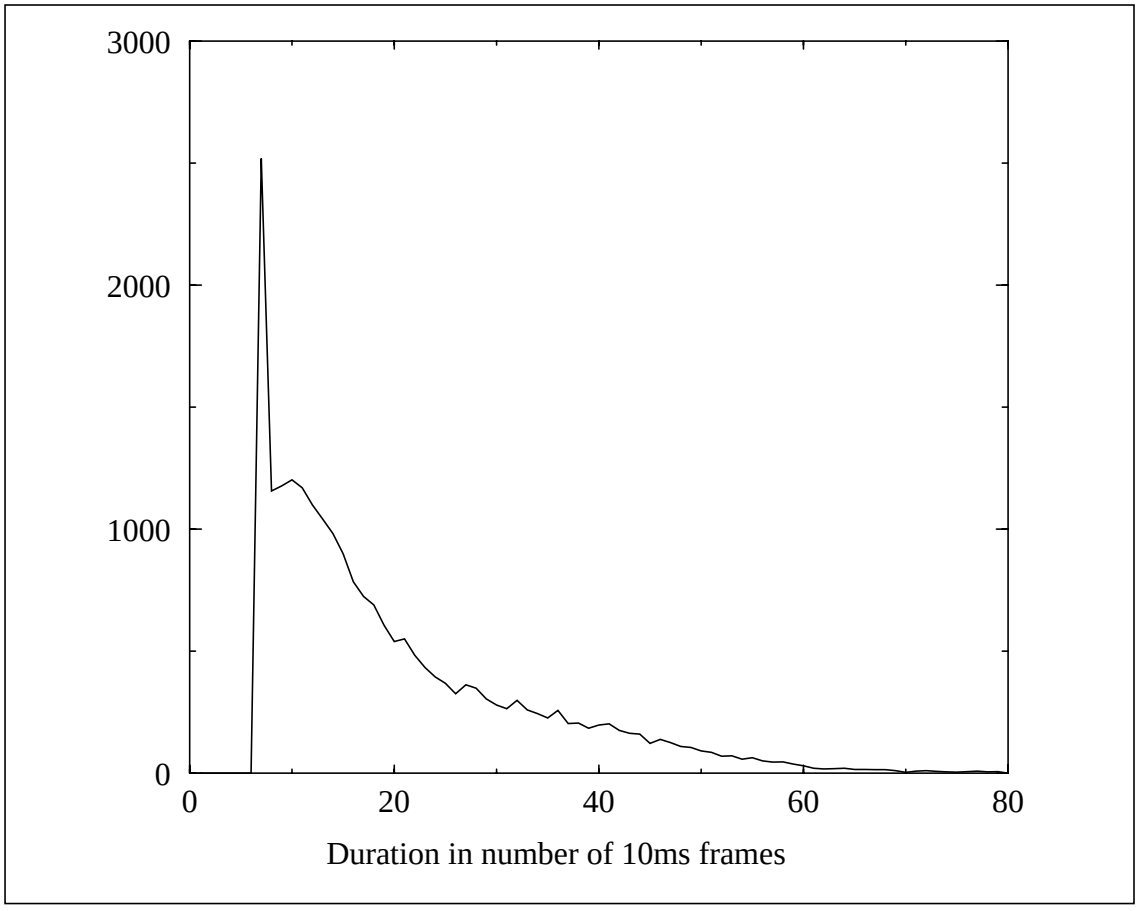

Figure 3: Duration histogram for the syllable “_ae_n_d”

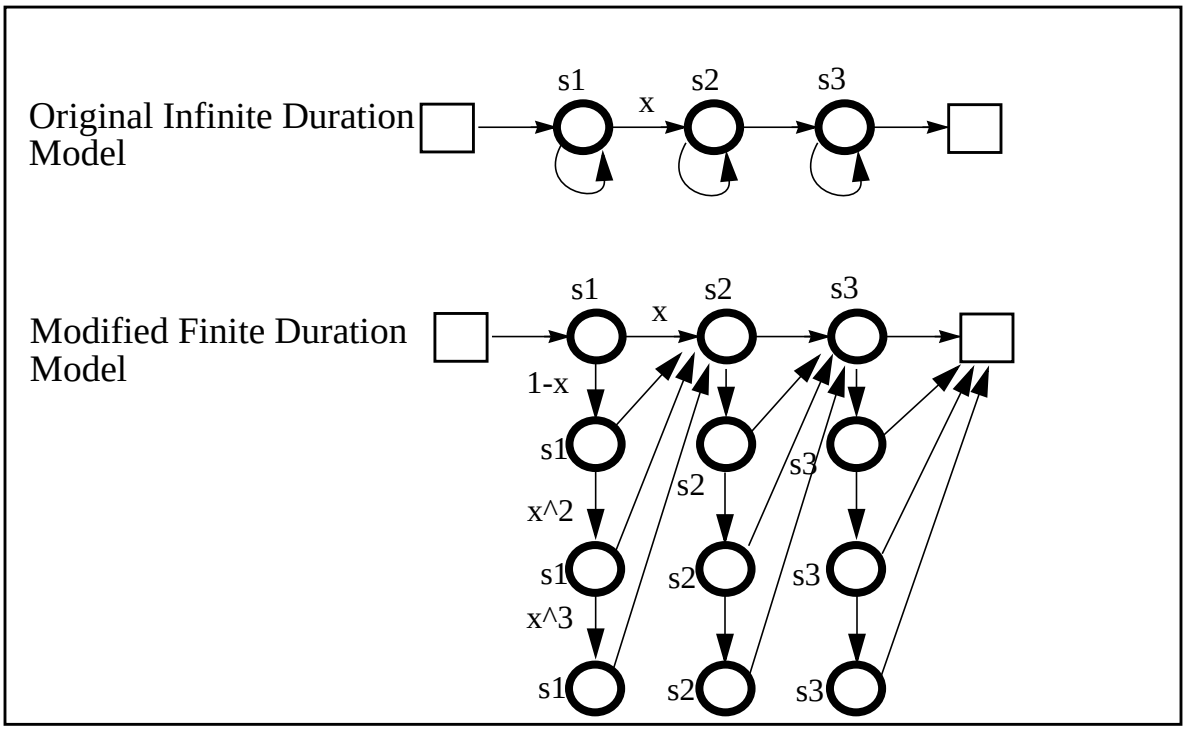

Figure 4: Topology modification, Top_001

could be classified into one of three categories: *syllable-only* (SO) words — words that have one or more syllables in their lexical representation but do not have any phones; *phones-only* (PO) words — words that have only phones in their lexical representation; and *mixed* (MX) words — words that are represented in terms of both phones and syllables. Table 1 shows a comparison of the errors for these two systems. The category '*miss*' represents incorrectly recognized or deleted reference words. The alignments required for this analysis come from the output of *sclite* [x], NIST's scoring software. It is evident from this analysis that the syllable system's performance degrades on MX and PO words, most likely attributable to edge effects at syllable-phone junctions. This was the motivation for our next experiment: a hybrid system with the context-independent phones replaced by their context-dependent counterparts.

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

ACCEPTED → _eh_k k-s+eh s-eh+p eh-p+t _t_ih_d

In this representation, the phone preceding the "-" specifies the left context, the symbol following the "+" specifies the right context, and any entity with an "_" is a syllable. 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. Figure 2 illustrates the process. This system achieved a WER of 51.7%. It is interesting to note that a system similar to this with context-independent phones resulted in an increased in the absolute WER of 4%. This highlights the significance of the edge-effect phenomena when syllables and phones are used

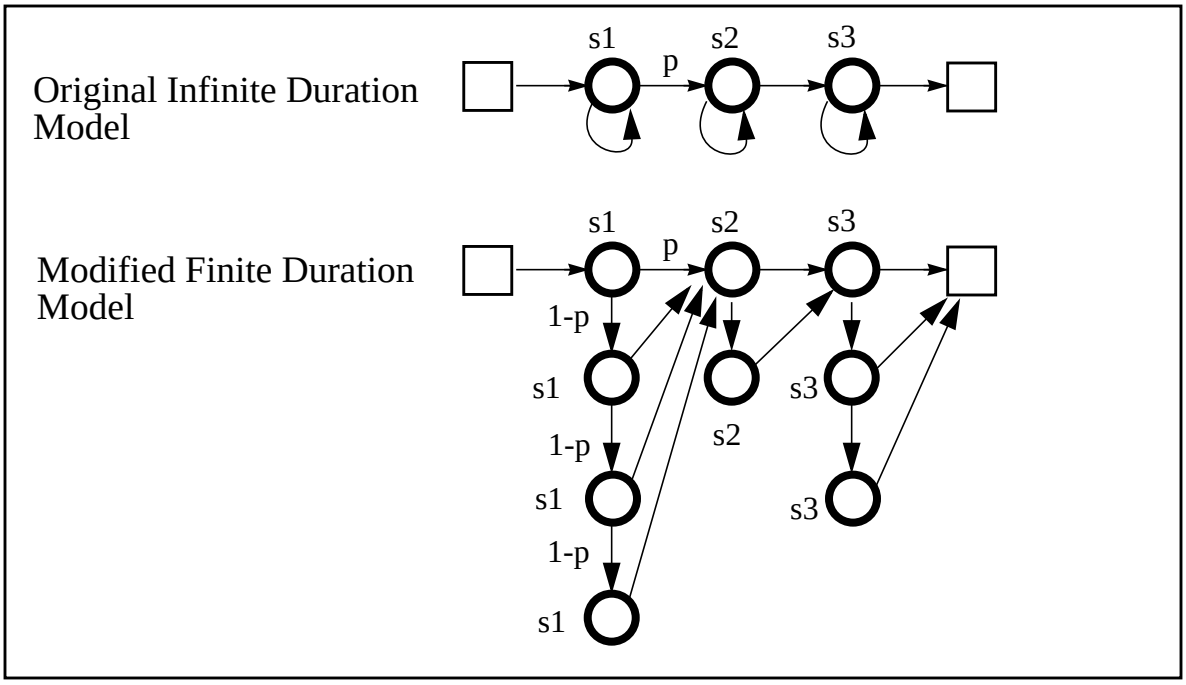

Figure 5: Topology modification, Top_002

together. Another important observation from the error analysis is that most of the errors (~70%) were on monosyllabic words. Modeling of monosyllabic words is a crucial obstacle to achieving high performance on SWB.

4. FINITE DURATION MODELING

The histograms of the syllable durations obtained from forced alignments show large amounts of variation from the nominal durations of the syllables. This is depicted in Figure 3. Further, the high word deletion rate with infinite duration syllable models, coupled with the expectation of higher durational stability from syllables suggests a need for additional durational constraints on the syllable models. Our first attempt at accommodating this involved using a finite duration model topology. During the workshop we experimented with two finite duration topologies. Since our hybrid systems had a mixture of acoustic units, we had the option of applying finite duration topology for all the models or just the syllable models.

4.1. Top_001

The first topology we experimented with was a model topology where we replicate each of the states in the infinite duration model four times and remove the self loop for that state. Figure 4 shows how are infinite duration models were transformed to finite duration models. Replicated states have the same output distribution as the original state. Two strategies were used for the transition probabilities. The first strategy involved assigning transition probabilities as shown in Figure 4, where x was varied from 0.0001 to 0.1. The WERs as a result of this modification are listed in Table 2. The second strategy involved matching the assigning transition probabilities in proportion with the self-loop probability of the original state in the infinite duration model. This

allows a model to exhibit similar transition properties as the original model, yet have additional durational constraints.

4.2. Top_002

In the second set of experiments, each state (S) in the infinite duration model was replicated P times where P is a function of the expected number of frames mapped into the state S for a given syllable token. Mathematically,

$$P = E[N] + 2.stddev(N) = f(p) \quad , \tag{1}$$

where N is the number of frames mapped to the state S and p is the self loop probability for the state S. Note that P is a function of p, the self loop probability. As in Top_001, the output distribution of each of the replicated states was tied to the distribution of the original state. The newly generated model set was seeded from the infinite duration models and was further trained with 4 passes of Baum-Welch re-estimation. Figure 5 describes the change from an infinite duration topology to a finite duration topology. This modification was applied only to the syllable models in the hybrid system.

Table 2 summarizes the results of the topology modification experiments. It can be seen that the topology Top_001, as it stands, is not effective in improving WER. This may be attributed to the lack of training data for the output distributions and transition probabilities, or to the inadequacy of the topology itself. Top_002 on the other hand performs well when only applied to the syllable models.

5. MONOSYLLABIC WORD MODELING

The motivation to create monosyllabic word models was an intermediate step towards exploring

Experiment	% WER
Baseline Triphones	49.8
Hybrid Syllable	51.7
Top_001: $x = 0.0001$	67.2
Top_001: $x = 0.001$	66.4
Top_001: $x = 0.01$	64.0
Top_001: $x = 0.1$	63.4
Top_001: x a function of self-loop probability	58.1
Top_001 applied to syllables only	57.8
Top_002	49.9

Table 2: Summary of results for experiments on finite duration HMM topology

durational clustering for the 800 syllables + word-internal triphone system. The objective was to distinguish a syllable in a monosyllabic word from the same syllable appearing in words with multiple syllables. Our conjecture was that the duration of a syllable in a monosyllabic word should have a different durational distribution than the same syllable in words that are polysyllabic. For example, in comparing the syllable _ay, _ax, _ih_n to the respective corresponding monosyllabic word 'I', 'A', 'IN', we might conclude that the syllable _ay (which appears in words that are polysyllabic) would have a different durational distribution than the same syllable used in a monosyllabic pronunciation. Therefore, we built a system that contained the 200 most frequent monosyllabic words, retained the syllables that had enough training tokens, and added the word-internal triphones with the intention of realigning the training data to obtain durational distributions for syllables and monosyllabic word models. Surprisingly, this system reduced the WER by 2.4%(absolute) when compared to the system with 800 syllables + word-internal triphones.

Some interesting facts about this system:

- performance exceeds a comparable word-internal triphone system by an absolute WER of 0.5%;
- the WER remains unchanged when the 200 monosyllabic words are added to the word-internal triphone system.

This second result provides us with some evidence that monosyllabic word models are not the only reason for the improvement of our system over a word-internal triphone system. This alleviates the concern that the word-internal triphone system was unfairly constrained, since 71% of the word tokens in training are the 200 monosyllabic words.

5.1. Training Procedure

There was no specific basis for choosing 200 word models, except that all 200 models had sufficient training data. We settled on 200 word models after reviewing the effect of these models on the training tokens of the 60+ hour training set. The first step in creating this system was to use the 800 syllables + word-internal triphones to align the 60+ hour training set. From this alignment we

Category	Count/Percentage
Unique Words	15,127
Number of Word Tokens	659,713
Number of Monosyllabic Words (dependent on lexicon/alignments)	529
Monosyllabic word tokens covered by the top 200 Monosyllabic words	95%
Word tokens covered by the 529 Monosyllabic words	75%
Word tokens covered by the top 200 Monosyllabic words	71%

Table 3: 60 hour training data breakup

created a list of monosyllabic words, sorted by their frequency of occurrence. From this list we picked the 200 most frequent unique words (some monosyllabic words at this stage of the game had multiple pronunciations).

The 200 word models were seeded with the most frequent syllable model for that word. Now we had 800 syllables and 200 word models. However, the number of training tokens to train 800 syllables was reduced due to the creation of 200 word models (some syllables only appeared as monosyllabic words). A threshold of 114 training tokens was used to determine the syllables that would be trained. This number comes from the number of training tokens that were available to train the 800th most frequent syllable in the baseline syllable system (the syllable system that was used to seed all of the system mentioned in this section). Using this threshold only 632 syllables survived. It turns out that 168 syllables did not have enough training material and these syllables were stripped done to their monophone representations.

The number of states in the syllable and word models were reestimated by relabeling the forced alignments using 632 syllables and 200 word models. The number of states for each new model was again chosen to represent one half the median duration. The models of this hybrid system (200 word models + 632 syllables + word-internal triphones) were reestimated by running 4 iterations of the Baum-Welch algorithm on the WS'97 official 60+ hour training set. The hybrid system was tested on the official WS'97 test set using the official HUB5E scoring software [x].

The triphone models were updated only if a minimum of 114 training tokens existed. The default for HTK is 1 training token. All previous systems used the default parameter. In a follow-on experiment we trained and tested another system using the default value. Interestingly, we found

Category	**Number of Models affected**	**Example of some of the models (format: _model_name(# states in new model))**
Adding 7 states	1	_aw(11)
Adding 5 states	2	_g_aa_sh(23) _hh_ae_f(13)
Adding 4 states	5	_w_ey_r(13) _f_ah_n#(11) _ae_n(8) _g_eh_dh#(15) _f_ah_n(15)
Adding 3 states	18	
Adding 2 states	52	
Adding 1 state	146	
Not Changing	218	
Removing 1 state	160	
Removing 2 states	27	_ih_sh#(6) _#r_ih_n_t(5) _m_y_uw_z#(9)
Removing 3 states	3	_ey_sh#(5) _m_ae_g#(8) _d_ih_d(6)

Table 4: Change in model duration for syllables

Category	Number of Models affected	Example of some of the models (format: __model_name(# states in new model))
Adding 7 states	1	__yeah(17)
Adding 6 states	3	__oh(12) __uh(14) __um(22)
Adding 5 states	3	__true(17) __wow(23) __no(13)
Adding 4 states	6	__yes(18) __own(10) __news(17) __here(12) __huh(13) __i_(8)
Adding 3 states	4	__too(11) __right(18) __t_(10) __sure(17)
Adding 2 states	19	
Adding 1 state	57	
Not Changing	83	
Removing 1 state	20	
Removing 2 states	3	__don_t(8) __want(10) __our(8)
Removing 3 states	1	__the(5)

Table 5: Change in model duration for monosyllabic words

that this change accounted for only 0.1% of the 2.4% drop in absolute WER.

5.2. A Look at the Training Data

Table 3 categorizes the official 60+ hour training set according to the number of syllables in each word. The training set has a total of 529 monosyllabic words. It turns out that these 529 monosyllabic words cover 75% of the total number of word tokens in the training set. The top 200 monosyllabic words cover 71% of the total number of word tokens in the training set. However, the 329 monosyllabic words that we did not model account for only 5% of the total number of monosyllabic words in the 60+ hour training set.

5.3. Probing the 2.4% drop in WER

In comparing the 800 syllable + word-internal triphones system with a system consisting of 200 words + 632 syllables + word-internal triphones, we find there is an absolute difference of 2.4% in WER. This was a surprising result, but the reason for the decrease in WER error can be explained by comparing the various differences between these systems. The lexicon was one major difference. However, it was not the only difference between the two systems. The number of states in the models was also modified, and the number of training tokens needed to update model parameters was increased to 114 tokens. As explained earlier, it has been established in a follow-on experiment that this deviation from the normal training procedure (where a default of 1 token is used) accounts for 0.1% of the 2.4% drop in WER (absolute).

The number of states for each model was computed by relabeling the forced alignments on the 60+ hour training set using the 800 syllables + word-internal triphone system. The relabeling process was used to change the alignment labels to include 200 word models and the 632 syllable models. Syllable models that did not have enough training tokens were reduced to their corresponding monophone representations. Table 4 provides some insight into how the number of states change from the seed syllable model to the syllable model used with the 200 word models. One common concern is the number of training tokens that were affected by changing the durations. Syllables that grew by more than 1 state account for 12% of the syllable training tokens and syllables that lost more than one state account for 9% of the syllable training tokens.

The same type of analysis can be done on the word models. The word models were seeded by the most frequent occurring syllable for that word. There were 14 monosyllabic words that had more than 1 pronunciation. Table 5 provides some insight into how the number of states change from the seed syllable model to the word model. Word models that grew by more than 1 state account for 26% of the monosyllabic word training tokens and word models that lost more than 1 state account for 4% of the monosyllabic word training tokens. From these statistics, one can conclude that word models were influenced more by the change in durations than syllable models.

The experiment to account for the reduction in WER due to the change in durations was not completed due to time constraints. However, other durational clustering experiments run on the syllables did not have a major impact on the overall performance. It is unclear exactly what the outcome would be since ~50% of the training tokens for words and syllables had models that changed in duration by more than 2 observation frames (the models we are using assume that each state will consume, on the average, 2 frames of observation).

Word	Possible variations in original lexicon
THE	_dh_ah _ah_iy _dh_ax
FOR	_f_er _f_ow_r
TO	_t_ax _t_uw

Table 6: Examples of multiple pronunciations being folded into one model

Syllable	Word
_n_ow	KNOW, NO
_d_ey_r	THERE, THEY'RE, THEIR
_t_ax _t_uw	TO TOO, TWO

Table 7: Examples of separate models for words with distinct lexical baseforms

5.4. Effect on the Lexicon

Another contributing factor to the 2.4% improvement in WER is the lexicon. One reason why SWB is a difficult corpus to recognize is the variability in pronunciations. The deletion rate for phones in SWB has been estimated to be ~12%, whereas the deletion rate for syllables is ~1% (based on human transcription projects). Since, the syllable is a longer acoustic unit compared to the phone, the need to explicitly provide pronunciations for all variants could be alleviated. Therefore, it should be possible for the syllable model to automatically consume the acoustical variation of the pronunciation of a word/syllable in the model parameters. If this argument holds true, then the creation of word models allows for the variation in pronunciation to be clustered into autonomous models.

In the creation of word models two changes occur. First, some words that had multiple pronunciations were now represented by one model. Table 6 provides some examples for this modification. Another example not provided in the table is the word '*AND*'. In the lexicon there is only one pronunciation: "*_ae_n_d*." However, in conversational speech the possible alternative pronunciations could be a deleted "*_ae*," or a deleted "*_d*," or a deleted "*_ae*" and "*_d*." Using the larger acoustic unit is a way of making the word model less dependent on the lexical realization, and the variation in pronunciation can be modeled by the data directly.

Second, some monosyllabic words with the same baseform were now modeled as different word models. Table 7 provides some examples of this modification. The first example is probably the most obvious example. It is easy to imagine that the word "*NO*" can be short or long, and it is more unlikely that the word "*KNOW*" has this characteristic. The number of states in the models are different. The word model "*KNOW*" was built with 9 states and the word model "*NO*" was

Syllables Affected (800 syllable system)	System that had 200 words + 632 syllables	Coverage of the 60+ hour training set (Excluding Phones)	What the syllable became
58	Monosyllabic Words (__and __that)	19%	Word Models
98	Monosyllabic Words (_it's _ih_t_s)	21%	Word Models
	Phones (graduates, diets, poets)	< 114 Tokens	Phones
12	Phones	< 114 Tokens	Phones
50	Monosyllabic Words (__i __a __in)	24%	Word Models
	Syllables (_ay _ax _ih_n)	7%	Syllables
582	Syllables	29%	Syllables

Table 8: Mapping 800 syllables into 200 word and 632 syllable models

Syllables Affected (800 syllable system)	System that had 200 words + 632 syllables	Misses	False Alarms
58	Monosyllabic Words (__and __that)	+0.1%	-2.0%
98	Monosyllabic Words (_it's _ih_t_s)	-0.1%	-2.7%
	Phones (graduates, diets, poets)		
12	Phones		
50	Monosyllabic Words (__i __a __in)	-1.8%	-3.0%
	Syllables (_ay _ax _ih_n)		
582	Syllables		

Table 9: Effect of mapping 800 syllables to 200 word and 632 syllable models

built with 13 states. The difference between these models is on average 80ms of speech. Similarly, the third example depicts a situation where both the number of states in the model as well as the most likely pronunciation of the word would vary. The word "*TO*" is more likely to be pronounced as "*_t_ax*" rather than "*_t_uw*" in conversational speech. In this case the number of states needed to model "*TO*" is four compared to 10 states for the word "*TWO*," and 11 states for the word "*TOO*."

5.5. Effect on Word Models

Though we have provided examples of the changes in the lexicon, we have not discussed the relationship between word models that used to be syllables and vice versa. These effects can be explored in several different ways. First, we can examine how syllables were affected by the creation of word models by analyzing the change in coverage. After the categories of change have been established we can determine how the errors changed for monosyllabic words in terms of misses (reference words that either deleted or substituted) and false alarms (hypothesized words that were either inserted or substituted).

One might wonder how the 800 syllable models in the baseline systems were changed to 832 models (200 word models and 632 syllables) in the monosyllabic word system? The reason is that, some of the syllables were trained only from monosyllabic word tokens and some of the syllables had training tokens that were both from monosyllabic and polysyllabic words. However, when word models were created, some of the original syllables ended up having insufficient training material to accurately train both a word model and a syllable model. For a model to have enough training tokens there needed to be 114 examples. This was based on the fact that the 800th most frequent syllable was trained on 114 tokens. Table 8 provides a breakdown on how the 800 syllables were mapped to the 200 words + 632 syllables. 58 of the 800 syllables were already

monosyllabic words and this covered 19% of the 800 syllable tokens in the 60+ hour training set. Note that some of the syllables had multiple pronunciations for a monosyllabic word and when we created a word model we took the most frequent syllable for that monosyllabic word to seed the model. Finally, 234 syllables were in fact monosyllabic words that cover 10% of the non-phone training tokens.

Now that we have established how the syllables changed into word models, Table 9 shows the breakdown of errors in terms of misses and false alarms. The misses and false alarms is the absolute change of error from the 800 syllables + word-internal triphone system to the 200 words + 632 syllables + word-internal triphones system. The change is computed for the words under each distinct category. For example, for 58 syllables that were already word models, the miss rate increased by 0.1% and the false alarm rate decreased by 2.0%.

5.6. Finite Duration Word Models

The finite durational models presented in this final report were implemented on the 200 word models and 632 syllables. The word-internal triphone models were not transformed to finite duration models (remained as 3 state left-right HMMs). The word error for this system on the official WS97 test set was 49.1%. This is compared to 49.3% WER for the 200 word models + 632 syllables + word-internal triphones using infinite duration models.

6. AUTOMATIC IDENTIFICATION OF MODALITIES

One of our goals in the summer workshop was to test the hypothesis that syllables allow one to carry out research that is difficult in phoneme based recognizers. Our method of demonstrating this was to attempt to develop a successful scheme for automatically identifying modalities in LVCSR acoustic data (which traditionally has been a difficult task with phone based recognizers).

What are modalities? Modalities are classifiable variations in acoustic data. Well known examples are pronunciation variations due to gender, dialect or context. A modality may also be produced by a significant fraction of some word tokens lacking a constituent phoneme (as in pronouncing "*and*" as "*n*"). The use of modalities is well established in speech recognition. We build gender dependent models and context dependent triphones. Explicitly modeling modalities produces sharper models. But these are heuristically defined modalities, not necessarily suited for all recognition tasks. Gender models, for example, are not useful for small corpora. Dividing a small corpus by gender produces badly trained models.

If we could however automatically detect the presence of modalities, we could reap the benefit of sharper models when the acoustic data supports their existence and not pay the penalty from under-training, when data does not contain the modality. The use of classification trees driven by linguistic questions during the training of context dependent triphones is an example of this strategy. In addition, automatic identification could also uncover modalities unknown to the researcher. For example, telephone data may have been collected on different handsets. In spite of the promise of automatic detection of modalities, no successful example of an implementation for this in LVCSR is known to us, although attempts have been made. We review these attempts, as well as successful implementations in the area of digit recognition in the next section.

6.1. Previous work

The effectiveness of automatic identification of modalities in continuous digit recognition has been demonstrated previously [11]. Whole word models for digit tokens were clustered using a dynamic programming approach that simultaneously considered spectral and duration information. This scheme could be easily applied to our syllable recognizer.

Another approach in this direction has been the parametrization of trajectories in acoustic space. Gish and Ng have developed a formalism for parametrizing trajectories and defining a distance metric between them [4]. Kannan and Ostendorf subsequently used this approach to cluster triphone models in a recognizer trained on SWB [12]. They obtained a small improvement of 0.1% over models using full Gaussians and no mixtures. This gain was lost once tying of models was allowed on both systems. Parametrizing syllable trajectories is certainly an appealing idea and we hope that it will receive more attention in the future.

Lastly, the work of Korkmazskiy et. al. also demonstrated a successful method of automatic identification of modalities in continuous digit recognition [6]. They developed a technique for identifying modalities in a wireless telephony corpus. Their novel approach was not to cluster the tokens for a digit by looking at their acoustic feature vectors but instead at their log scores from forced alignments. They then trained an HMM on each cluster and combined them as separate paths in a multipath HMM for the digit.

A pessimistic assessment of these three approaches to automatic identification would state that the success of automatic methods in digit recognition is due to the relative ease of the task in comparison to LVCSR. An optimistic one would state that the success was due to the larger unit of speech used in digit recognition. We believe that the longer duration of syllables makes them suitable for the development of automatic identification techniques in LVCSR.

6.2. Experimental Design

Three main experiments were performed to identify modalities. We began with a direct implementation of the ideas of Korkmazskiy et. al. in our syllable recognizer. Our second experiment implemented the dynamic programming technique. And lastly we investigated a new approach based on identifying preferred Gaussian transitions between HMM states. Common to the design of all our experiment was the use of multipath HMMs. Implementing modalities as pronunciation variations in the dictionary would have intersected with this syllabification work. Instead, we chose to keep these two areas of work separate by implementing modalities as multiple paths within a single HMM, a straightforward procedure in HTK based training and testing.

6.3. State Space Clustering

Research on acoustic syllable models at this workshop was limited to context-independent models. It is, however, widely known that context-independent acoustic models have only a limited capacity of accounting for the acoustic variation in the speech signal. First, lack of context information prevents them from modeling co-articulation phenomena. Second, their modeling accuracy is usually impaired by more general variation caused by different speakers, sex, dialects, recording conditions etc. In order to improve the performance of our syllable models, we

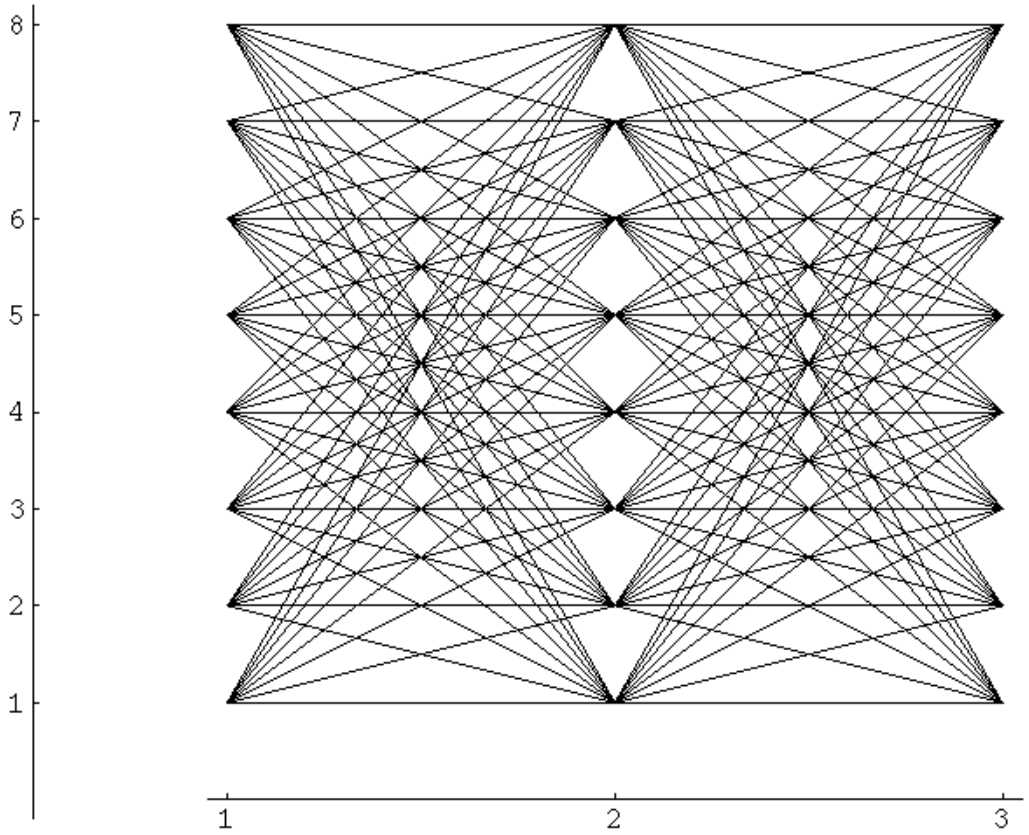

Figure 6: State mixture space; 3 states and 8 mixtures giving 8^3 paths

investigated techniques other than context-dependent modeling, which were designed to remedy certain inadequacies in HMMs leading to weak modeling power.

A major shortcoming of standard continuous density HMMs is the amount of non-determinism inherent in their topologies. In a fully continuous HMM the emission probability of an observation vector at any time instant from a given state is computed as a weighted sum ("mixture") of m Gaussian probability density functions (or "mixture components") associated with that state. In practice, however, the emission probability in a particular state is clearly dominated by a small set of mixture components, in the sense that one or two mixture components receive high scores and contribute most to the overall sum, whereas the remaining mixture components receive low scores and have little impact on the resulting sum. A typical decomposition of a mixture is shown in Table 10.

Also, the selection of dominant mixture components during decoding is unconstrained in the sense that it only depends on the acoustic vector at the current time frame. It is independent of the choice of mixture components at the previous and/or following time frames, by virtue of the Markov property of HMMs. Thus, the combinatorial possibilities of dominant mixture components across states can be described by a directed acyclic graph defining the state-mixture space (Figure 6). Nodes in this graph stand for mixture components. An HMM with n states defines $n-1$ complete subgraphs: all pairs of adjacent nodes are fully connected. For n states and m mixture components per state, the number of possible paths through this graph is m^n.

Mixture Component	Weight (log)	Acoustic Score (log)	Weighted Score (log)
state 2			
1	-2.554125	-113.714073	-116.268198
2	-2.009913	-83.053894	-85.063807
3	-1.562939	-68.474358	-70.037297
4	-2.267671	-104.138321	-106.405992
5	-1.909643	-83.689354	-85.598997
6	-2.051309	-78.683815	-80.735124
7	-2.269707	-101.681313	-103.951020
8	-2.352568	-106.311897	-108.664465
sum: -70.037274			
state 3			
1	-2.380755	-79.351425	-81.732180
2	-2.030592	-73.052673	-75.083265
3	-2.168982	-78.110313	-80.279295
4	-2.032470	-79.761726	-81.794197
5	-1.771929	-64.594414	-66.366343
6	-2.062797	-76.163620	-78.226417
7	-2.451171	-73.986580	-76.437750
8	-1.912837	-69.708214	-71.621050
sum: -66.360920			
state 4			
1	-3.252082	-153.424118	-156.676200
2	-1.901964	-87.951492	-89.853456
3	-1.774892	-82.281158	-84.056051
4	-2.238418	-73.746498	-75.984916
5	-1.716696	-73.561981	-75.278678
6	-1.952774	-85.651733	-87.604508
7	-1.762484	-82.664116	-84.426600
8	-3.152487	-102.110382	-105.262869
sum: -74.877380			

Table 10: Mixture decomposition, syllable _k_uh_d. The dominant mixture components are those with the largest weights

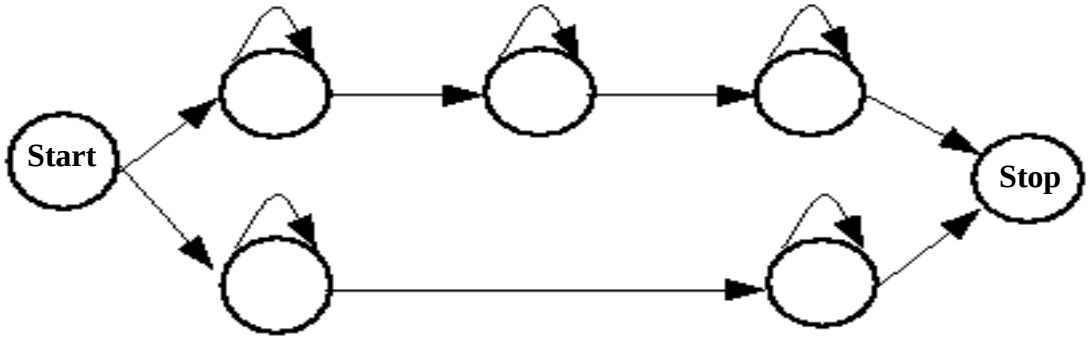

Figure 7: Multipath HMM structure

```
0 1400000 s2 -64.901794 sil -65.584816 [SILENCE]
1400000 1700000 s3 -64.778069
1700000 1800000 s4 -77.567467
1800000 2000000 s2 -80.565437 _ow -78.388321 OKAY
2000000 2300000 s3 -82.814980
2300000 2400000 s4 -70.224297
2400000 2500000 s5 -69.184402
2500000 2700000 s6 -74.257866
2700000 2800000 s7 -86.383003
2800000 3000000 s2 -85.982735 _k_cy -73.132599
3000000 3200000 s3 -86.620911
3200000 3400000 s4 -80.245056
3400000 3600000 s5 -75.816124
3600000 4100000 s6 -78.143051
4100000 4200000 s7 -65.643066
```

Figure 8: State level alignment used for state space clustering

In practice, however, some mixture components group together across states to model particular modalities. These horizontal clusters, which we call mixture trajectories, are not enforced by the HMM topology; rather, continuity information is typically lost due to the Markov assumption. This may lead to crossovers of different trajectories, which aggravates confusion between different models during decoding [14]. However, if the number of possible trajectories could be limited, confusion would be reduced.

Our goal was to apply an algorithm to syllable models which would automatically separate relevant trajectories and create model structures which explicitly disallow crossovers between them. One such model structure is a multipath model, which assigns trajectories to separate unconnected paths in an HMM as shown in Figure 7.

The main research issues in this paradigm were: the data-driven extraction of trajectories through the state-mixture space, and the optimal topologies for the resulting HMMs. Two different methodologies for acquiring trajectories were investigated: implicit acquisition of trajectories by clustering the training data and retraining new models for each of the cluster, and explicit extraction of mixture trajectories from original models.

In order to set up the experimental environment for investigating multipath models, a pilot experiment was conducted which used clustering of the training data as a basis for determining mixture trajectories. The training data was clustered according to a distance measure proposed in the literature, viz.~state acoustic likelihood [6]. The state acoustic likelihood is the likelihood of the data given a model, averaged over all frames assigned to a particular state. These values were

obtained by a forced alignment of the entire WS97 training set using the hybrid syllable system's models (800 syllables and word-internal triphones). An example of this state-level alignment is shown in Figure 8.

Since the models used for alignments did not contain skip transitions, each instance of a model could thus be assigned a fixed-dimensional vector, where the number of vector components was the number of states in the model. These vectors would then serve as the basis for splitting the data into two clusters, one cluster comprising those examples which are distributed in the high probability region and thus match the model well, and a second cluster which subsumes tokens in the low probability region which exhibit a greater distance from the original model. The data was split along the mean of the distance vector distribution, i.e. along the plane perpendicular to the direction of greatest variation. This split is shown in Figure 9 for a three-state model.

The splitting procedure was applied to the 485 most frequent syllables, out of 800 syllable models. Each of these models had more than 200 training tokens, which was considered the minimum threshold in view of the fact that fewer than 100 tokens would remain for training after data splitting. Each token in the training database was relabelled as either belonging to Cluster 1 or to Cluster 2. For each cluster, a separate model was subsequently trained. The topologies of these cluster models were determined as follows:

- manually created a new segmentation of the training database that consists of utterances typically 10 seconds in duration and are excised at significant pause boundaries and/or turn boundaries;
- the average (mean) duration of the tokens in each cluster was recomputed;
- based on the mean duration, the number of states in each model was chosen;
- as in the initial procedure, models were designed to contain one state for every two frames;
- the final number of mixture components per state was defined as half the number of mixture component per state in the original model.

The clustered models were subject to an iterative process of splitting the Gaussians with four subsequent passes of re-estimation, until four mixture components per state were obtained. In a final step, the separate cluster models were merged into a single model by connecting them to identical START and STOP states. These cluster models were then used for decoding, together with the 315 remaining unclustered syllable models and triphone models. It should be noted that this way of splitting the training data was not designed to directly identify mixture trajectories. It was expected that the clustering and re-training procedure would automatically separate mixture trajectories.

Decoding was carried out on the WS'97 development test set. Compared to the baseline syllable system, the word error rate dropped by 0.2%. A more detailed error analysis showed, however, that the syllable models did not improve as expected. The number of misses for all-syllable words (this includes monosyllabic words and thus the most frequent syllable models) increased from 46.3% to 47.4%. On the other hand, words consisting of both syllable and phone models as well as those consisting only of phone models did improve substantially as far as the number of misses is concerned, though, the number of false alarms increased. These results seem to indicate that

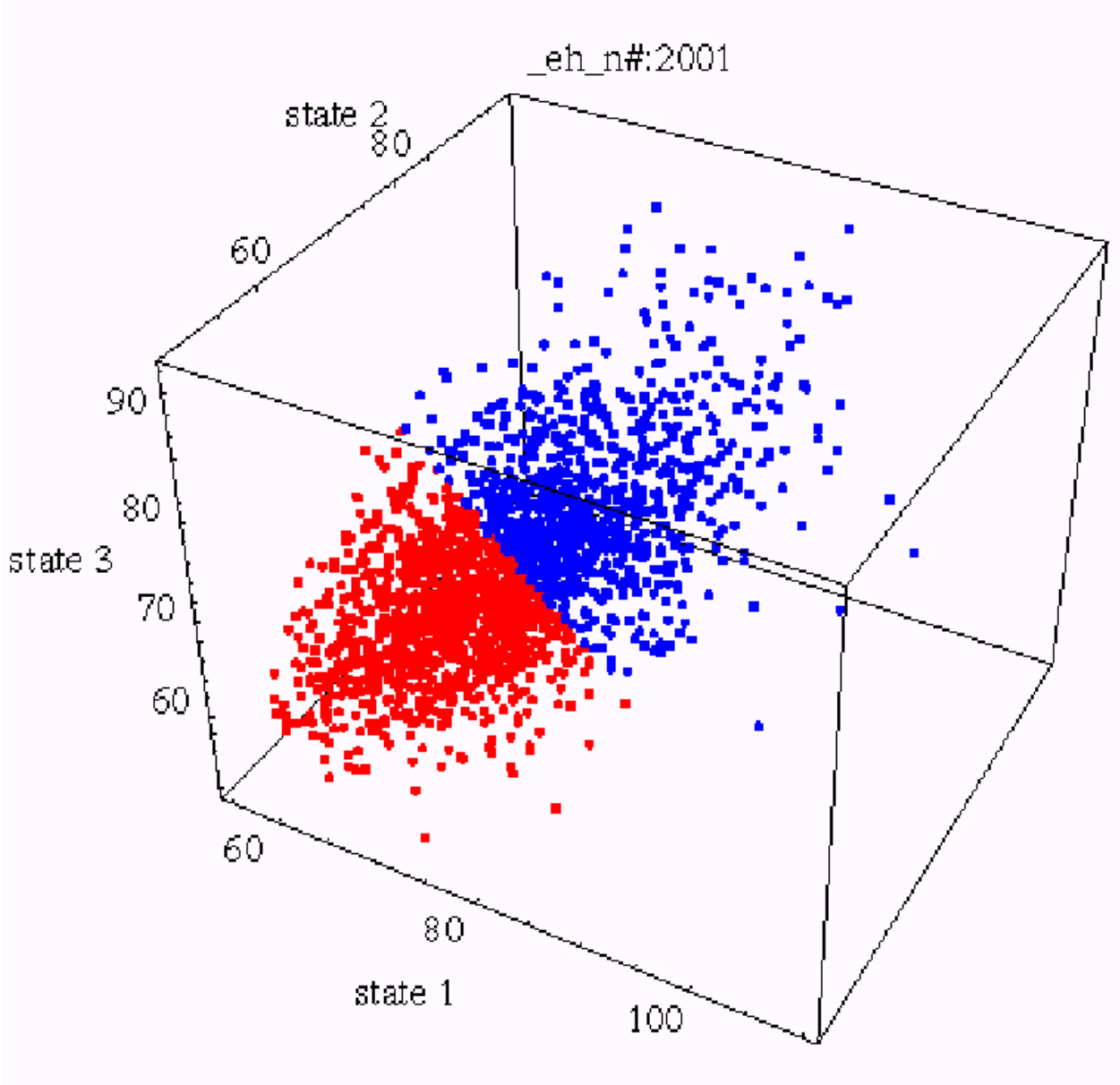

Figure 9: State space clustering, two clusters in this case

though syllable models deteriorated, the clustering procedure lead to a better treatment of words consisting either of both syllables and phones or of phones only. Therefore, the marginal improvement in word error rate might be due to the better acoustic modeling at syllable-triphone boundaries. It had been observed before that edge effects at these boundaries have a significant impact on the overall system performance.

6.4. Durational Clustering

It was observed during the course of the workshop that many syllables exhibit a durational distribution with a large variance, as shown in Figure 3. This led to the assumption that a more accurate modeling of various durational modalities might improve performance. We therefore

decided to create multipath models with each path in the model representing a durational modality. The models were created as follows:

- first, a forced alignment was generated on the training set using the syllable system which comprised 200 monosyllabic word models, 632 syllable models and triphones;
- based on those alignments, duration histograms were computed for each of the monosyllabic word models and for the top 260 syllable models;
- the data was split at the 25th and 75th percentiles of the duration histogram for each model;
- the numbers of states for the multipath models were re-defined: all tokens below the geometric mean were assigned a number of states equivalent to half the duration at the 25th percentile, all tokens above the geometric mean were assigned a number corresponding to half the duration at the 75th percentile;
- the new models were seeded from the original models. Where the new model had fewer states, states were deleted from the beginning/end of the original model; in cases where the new model contained more states, states from the original model were replicated and inserted at equal intervals in the new model;
- finally, four passes of re-estimation were performed on the new models.

Decoding was carried out using the multipath models, the remaining 372 syllable models and the triphone models. This strategy however increased the word error rate by 0.2%. The numbers of both false alarms and misses increased for all word types. This may be due to the fact that the durational characteristics have been constrained excessively by allowing only two basic duration modalities for each multipath model. Furthermore, the new models were not re-trained from scratch but seeded from the original models, which clearly is a severe limitation.

6.5. Dynamic Programming Clustering

The basic idea behind the dynamic programming approach to identify modalities is to calculate the distance between training tokens and then apply the K-MEANS algorithm to obtain clusters. Our implementation began by mapping each frame of a token for a speech unit to a Gaussian and state label. The state labels were obtained from a forced alignment of the training data at the HMM state level. The Gaussian label for a frame corresponded to the most probable Gaussian from the state mixture model.

The distance between tokens was calculated by a dynamic programming scheme. Consider two tokens, possibly of different durations. Begin by stretching the shorter one to align with the longer token. This involves some arbitrary assignment of Gaussian and state labels to the new “frames” of the shorter token. You can now compare both tokens on a frame by frame basis. If the state labels for a frame differ between the two tokens, you incur a penalty of two. If the state labels are the same for a frame but the Gaussian labels differ, you incur a penalty of one. The distance between the tokens is the sum of these frame penalties. But there are other ways of assigning labels to the stretched token. We use the assignments that minimizes the penalty between them. As defined, however, this distance is not symmetric. So another minimizing distance is calculated by shrinking the longer token to the duration of the shorter one. The final distance used is the

average of the two minimizing distances. Using this distance metric, clusters are identified.

6.6. Gaussian Transitions Clustering

It is commonly stated that speech violates the assumptions inherent in the use of HMMs. One such assumption is that the speech frames of different HMM states are uncorrelated. During the workshop, we formulated a simple procedure to test this hypothesis for the case of mixture models.

Instead of looking at acoustic vectors, one can consider the Gaussians in a mixture model. When we train mixture models, we are assuming that the Gaussians in different states are conditionally independent. If we assign the frames of a token to particular Gaussians in a state, we would expect those assignments to be statistically independent of similar assignments for frames assigned to another state. The existence of modalities in our training tokens would violate that assumption. We can quantify this violation by implementing a variation of the procedure used by Korkmazskiy et al.

Note that, from the state space clustering approach all tokens had been reduced to the same length by considering the average score for the frames assigned to a state of the HMM. Thus, all tokens of a three-state HMM could be mapped to a three dimensional space irrespective of their differences in total frame lengths. Since we were searching for modalities in the data we began to study the sequence of Gaussians produced by a token as it traversed the HMM. To that end, we applied the following ad-hoc procedure using the forced alignments done with our best syllable recognizer.

Each frame in a state of a training token was assigned its highest scoring Gaussian from the mixture components for that state. The Gaussian index, along with its score, formed a data pair for each frame of a token: {highest scoring Gaussian index, its score}. These pairs were then organized as follows:

state data list = {one or more frame pairs assigned to a state}

token data list = {one or more state data lists}

These data lists were then reduced to a length equal to the number of states in the HMM by mapping each state data list to the index of the most common Gaussian in each state. If two or more Gaussians had the same largest frequency, we broke ties by selecting the highest scoring Gaussian. This reduction maps all tokens to a vector of the form {g(1), g(2),...,g(n)}, where g(i) denotes the Gaussian index for state i. We call such a vector a Gaussian sequence.

This procedure may seem crude and others may be designed but for the purposes of our investigation it has a crucial virtue. It does not introduce statistical dependence between frames assigned to different states of an HMM. Thus, any test of the null hypothesis that there is no dependence between Gaussians from different states will still be valid.

Initially, we had hoped that a frequency count of the Gaussian index vectors would reveal dominant “trajectories”, in the tokens. A look at the data for “__a” quickly showed that this was

not a useful approach. The monosyllabic word "__a" was the most common speech unit in our training data. It had 15,316 training tokens. Its HMM had three states with 8 Gaussians per state. Thus, we expected that the table of frequencies for the Gaussian sequences should have an entropy of roughly 9 bits (8^^3) given that the Gaussians had roughly the same weight in the mixture model. The observed frequency, however, only required about 7 bits.

We concluded from this that there was evidence for modalities in the training tokens. Unfortunately, the most commonly observed Gaussian sequence corresponded to less than 5% of the tokens. Creating a separate path in the HMM for such a small percentage of the tokens seemed a fruitless approach. Most of the syllables have a frequency of one or two thousand tokens and 5% of that, about a 100 tokens, is marginally sufficient for reliable estimates of the Gaussian parameters. In addition, most syllables have more than three states so whatever "trajectories" we identified by this procedure would rapidly blend in with the statistical noise. Instead, we chose to look at the correlations between Gaussians in different states.

If "trajectories" in the Gaussian sequences were not common enough, perhaps transitions between states would be. In the case of the "__a" tokens, we found that 9% of the tokens made a transition from Gaussian 4 in state 2 to Gaussian 6 in state 3. This frequency was about six times higher than expected from the mixture weights assigned to those Gaussians. This phenomena was also present in other syllables. For example, the monosyllabic word "__or" had a six-state HMM and 3,330 training tokens. About 19% of the tokens went from Gaussian 8 in state 3 to Gaussian 4 in state 4. Due to time constraints we could not explore this approach further.

There are various ways of quantifying the violation of statistical independence in the transitions observed in the Gaussian sequences. One can look at the Kullback-Liebler distance between the assumption of conditional independence and the observed distribution [13]. Alternatively, a chi-squared test on the contingency table built from the frequencies of the observed transitions between states could be calculated. In either case, one would have a measure of what transition seems to be most correlated. For example, the transition between states 3 and 4 in the "__or" example above. The training tokens could then be clustered so as to minimize transition correlations within clusters. Multipath HMMs could be trained in a manner similar to the approach outlined in our previous experiments.

7. SUMMARY

In this 1997 Summer Workshop on Innovative Techniques for LVCSR, the aim of the Syllable Speech Processing Team before the workshop began was to study the feasibility of using syllables as units for acoustic modeling for LVCSR and to compare its performance with state-of-the-art triphone based systems. Various baseline systems were built, namely, a word-internal triphone system, a cross-word triphone system, a context independent phone system and an all syllable system. Due to various reasons such as decoder availability, and ease of analysis, we decided not to experiment on context-dependent syllables during the workshop. As a fair comparison we therefore compare our syllable systems with the word-internal triphone system.

One of the major contributions of our group was developing a strategy to efficiently train and test systems which use models covering a wide range of temporal/linguistic contexts like words,

System Description	% WER
Context Independent Phone System	62.3
Context Dependent Cross-Word Triphone System	45.6
Context Dependent Word-Internal Triphone System	49.8
Baseline Syllable System, 800 Syllables and 42 Monophones	55.1
Hybrid Syllable System, 800 Syllables and Word-Internal Triphones	51.7
Hybrid Syllable System, 200 Monosyllabic Words, 632 Syllables, word-internal Triphones	49.3
Finite Duration Monosyllabic Word System	**49.1**
State-Space Clustering on Hybrid Syllable System with 800 Syllables	51.5

Table 11: Summary of significant results

syllables and phones. Table 11 summarizes results of the significant experiments. (the group finally ended up with a whopping 50 experiments with at least 4 passes of re-estimation in each case).

A close look at the SWB data shows that over 70% of the database consists of monosyllabic words. Most errors in the hybrid syllable system were on monosyllabic words too. This prompted us to model the monosyllabic words explicitly as word models. We chose to model 200 monosyllabic words which covered over 75% of the monosyllabic word instances in the training data. This system comprising of monosyllabic words, syllables and word-internal triphones performed at 49.3% WER surpassing the word-internal triphone system by 0.5% absolute WER. Applying finite duration topology to the models reduced the WER further to 49.1%.

Another area of work conducted during the summer was to automatically model modalities. One such strategy was the state-space clustering technique using multipath HMMs. This system gave a marginal improvement of 0.2% over the baseline. Other techniques like Gaussian transition clustering were trained but could not be tested due to time constraints.

The results from this summer's work clearly indicate that syllables show promise in modeling acoustics for LVCSR. In post-workshop research conducted at Mississippi State University syllables were used on a smaller domain, AlphaDigits [15]. The syllable based system trained and tested similar to the SWB systems outperformed a cross-word triphone system by a 2.5% absolute or a 20% relative WER. This clearly validates the results achieved at this workshop. In another post-workshop experiment a system comprising of word-internal triphones and 200 monosyllabic words was trained and tested and not surprisingly this system did not improve performance. This shows that the improvement in performance of the syllable system over its baseline is not attributed only to modeling monosyllabic words explicitly. There exists a more subtle advantage in modeling syllable-sized units for LVCSR.

As part of future work we need to understand the effect of ambisyllabics in the lexicon. For expediency we ignored issues involving the ambisyllabics during this summer's work. More efficient modeling of suffixes and common word endings also stands for further research and exploration. Note that none of the syllable-based systems described in this report used state-tying to tackle under-trained models. This is another area that needs further investigation.

REFERENCES

[1] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 517-520, San Francisco, California, USA, March 1992.

[2] H. Bourlard, H. Hermansky and N. Morgan, "Copernicus and the ASR Challenge -- Waiting for Kepler," *Proceedings of the DARPA Speech Recognition Workshop*, pp. 157-162, Harriman, New York, USA, February 1996.

[3] S. Greenberg, "The Switchboard Transcription Project", presented at the *1996 LVCSR Summer Research Workshop*, Johns Hopkins University, Baltimore, Maryland, USA, August 1996.

[4] H. Gish, and K. Ng, "Parameter Trajectory Models for Speech Recognition," *Proceedings of the IEEE International Conference on Speech and Language Processing*, pp. 466-469, Philadelphia, Pennsylvania, USA, October 1996.

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

[6] F. Korkmazskiy, "Generalized Mixture of HMMs for Continuous Speech Recognition," *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 1443-1446, Munich, Germany, April 1997.

[7] A. Ganapathiraju, et. al., "Syllable - A Promising Recognition Unit for LVCSR," To appear in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, California, December 1997.

[8] M. Ostendorf, et. al., "Modeling Systematic Variations in Pronunciations via a Language-Dependent Hidden Speaking Mode," presented at the *1996 LVCSR Summer Research Workshop*, Johns Hopkins University, Baltimore, Maryland, USA, August 1996.

[9] D. Kahn, *Syllable-Based Generalizations in English Phonology*, Indiana University Linguistics Club, 1976.

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

[11] J. Picone, "Duration in Context for Speech Recognition," *Speech Communication*, vol. 9,

no. 2, pp. 119-128, April 1990.

[12] A. Kannan and M. Ostendorf, “Adaptation for Acoustic Modeling at Boston University,”, *Proceedings of the LVCSR Workshop*, May 1997.

[13] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*, New York, NY: John Wiley & Sons, 1990.

[14] I. Illina and Y. Gong, “Elimination of Trajectory Folding Phenomenon: HMM, Trajectory Mixture HMM and Mixture Stochastic Trajectory Model,” *of the IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 1395-1398, Munich, Germany, April 1997.

[15] J. Hamaker, A. Ganapathiraju, J. Picone, and J. Godfrey, “Advances in Alphadigit Recognition Using Syllables” submitted to the *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Seattle, Washington, USA, May 1998.