A FOUR LAYER SHARING HMM SYSTEM FOR VERY LARGE VOCABULARY ISOLATED WORD RECOGNITION

Ruxin Chen, Miyuki Tanaka,
Duanpei Wu, Lex Olorenshaw, and Mariscela Amador
SONY Research Labs, 3300 Zanker Road, SJ2D4, San Jose, CA 95134, USA
TEL: 408-955-5374, FAX: 408-955-6530,
EMAIL: ruxin@lsi.sel.sony.com, http://kiku.lsi.sel.sony.com/ruxin

ABSTRACT

This paper reports on a large vocabulary speaker independent isolated word recognizer targeting 50,000 words. The system supports a unique four-layer sharing structure for either continuous HMM or discrete HMM. Evaluation is performed using a dictionary of 5000 US city names, a dictionary of the 5000 English most frequent words, a dictionary of 50,000 English words, and the 110,000 word CMU English dictionary. For these dictionaries, recognition accuracy ranges from 90% to 93% for the top 3 results.

1. HMM SPEECH RECOGNITION ENGINE

The speech signal is a one-dimensional waveform as shown in FIGURE 1. The speech signal may be labeled with a sequence of phonemes. A word may correspond to one or more continuous phonemes. An example of the phoneme labels of the isolated word "*item*" is also shown in FIGURE 1.

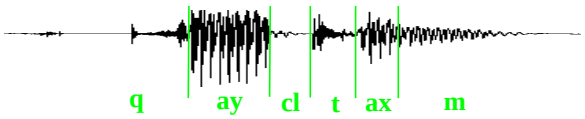

FIGURE 1 : Speech waveform and phonemes of the isolated word "*item*".

A left-to-right HMM process is used to model the speech waveform in this Speech Recognition Engine (SRE) as shown in FIGURE 2. This figure displays a simple 3-state left-to-right HMM of a phoneme where the context includes the left and the right phone, i.e., the HMM is context dependent. This type of HMM is chosen because it offers convenient flexibility for state sharing between the first, second, and last state of the HMM, as explained below. A series of HMMs correspond to a series of phonemes. The observations are emitted from each state of the HMM process. The observation probabilities can be formulated as probability distribution b_s, where s refers to a state in a HMM. Each state transition, shown as an arc in FIGURE 2, is associated with a state transition probability a_{sj} which denotes the probability of transitioning using the arc j of state s.

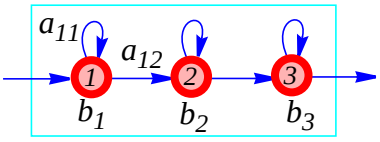

FIGURE 2 : A three state left-to-right HMM used in SRE for all the phonemes in a particular context.

Suppose there are B types of observations, and b_{si} denotes the distribution of state s and type (or stream) i, then

$$b_s = \sum_i b_{si}, \quad i = 1 \ldots B \tag{1}$$

The observations can be handled in two ways to create either a discrete observation HMM (DHMM) or a continuous observation HMM (CHMM).

The probability distribution of the discrete observation HMM defined as

$$b_{si} = b_{si}[q] = \sum_k b_{sik}[q], \quad q = 1 \ldots Q \tag{2}$$

is a one-dimension array with each scalar $b_{si}[q]$ denoting the probability of observing the vector quantized symbol q for state s, and $b_{sik}[\ ]$ denoting the sub-probability distribution that is a component in $b_{si}[\ ]$. Q in the equation denotes the total number of q. The sub-probability distribution b_{sik} is introduced for DHMM in order to compress the DHMM parameters more accurately and to share the structure between DHMM and CHMM more efficiently.

The probability distribution of the continuous observation HMM is defined as

$$b_{si} = b_{si}(o) = \sum_k b_{sik}(o)$$
$$b_{sik}(o) = \frac{c_{sik}}{\sqrt{|v_{sik}|}} \times \exp\left(-0.5 \times \frac{(o - m_{sik})^2}{v_{sik}}\right) \tag{3}$$

The diagonal Gaussian mixture is used to represent the probability of the continuous observation vector o for state s. c_{sik} is the weight for mixture k of state s and type i. Similarly m_{sik} is the mean for the Gaussian of mixture k; v_{sik} is the variance for mixture k.

To make the terminology convenient for both continuous HMM and discrete HMM, we will call b_{si} (in

EQUATION 2,3) the mixture-probability distribution (MPDTR), and b_{sik} the sub-probability distribution (SPDTR).

The novel EQUATION 2 has the same expansion formula as EQUATION 3. EQUATION 2 allows us to construct a common four-layer sharing structure for both discrete and continuous HMM. The interchangeable sharing structure for CHMM and DHMM make our speech engine differ from other systems reported in the literature[1,2,3].

2. HMM PARAMETER SHARING

The HMM parameter sharing consists of four layers. The first layer is the phoneme model sharing. The phoneme model sharing for the context dependent HMM as used by this system is exemplified by the HMM /w-ah+dx/. In this example, the HMM is a triphone, the main phoneme is /ah/ , /w/ is the left context phoneme, and /dx/ is the right context phoneme. In order to reduce the memory and to improve the robustness of the estimated model parameters, some HMMs may share similar portions of the triphones. The sharing of HMMs is illustrated in FIGURE 3. In the results from our preliminary experiments reported in this paper, we used the available corresponding data for each model during training to help determine the model sharing structure.

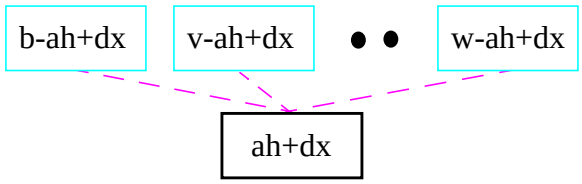

FIGURE 3 : Sharing of several triphone HMMs to a single biphone HMM model.

The second level of parameter sharing is the sharing among the states of different HMMs. When a group of states are shared, they have the same state transition probability a_s and the same observation distribution b_s. The sharing of states is illustrated in FIGURE 4.

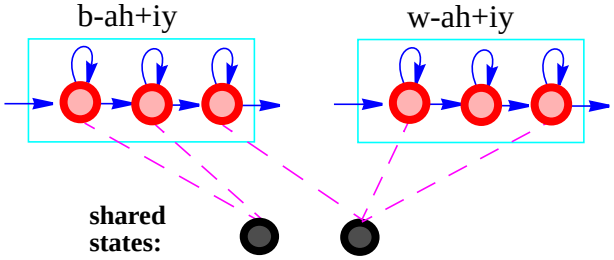

FIGURE 4 : Sharing of states of different HMMs.

The third level of parameter sharing is performed on the probability distribution b_{si}. This is illustrated in FIGURE 5.

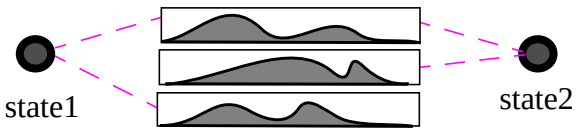

FIGURE 5 : Sharing of mixture probability distributions (MPDTR) between states. Each striped curve represents one distribution b_{si}.

FIGURE 6 displays the sub-probability distribution sharing. For continuous observation HMMs using Gaussian functions as SPDTRs, as shown, are shared with each other. On the other hand, for discrete HMM, FIGURE 6 illustrates the sharing of the SPDTR $b_{sik}[q]$.

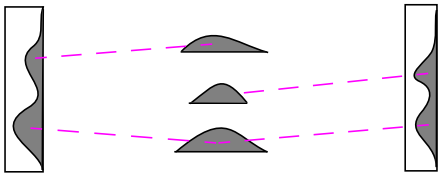

FIGURE 6 : Sharing of sub-probability distributions (SPDTR) among different probability distributions.

FIGURE 7 shows the overall four layer sharing structure of the Speech Recognition Engine(SRE). A top-down sharing design is usually performed first, that is, from models to states, from states to MPDTRs, from MPDTRs to SPDTRs. After the SRE parameters have been trained with speech data, a bottom up sharing procedure may be performed, that is, from SPDTRs to MPDTRs, from MPDTRs to states, and from states to models.

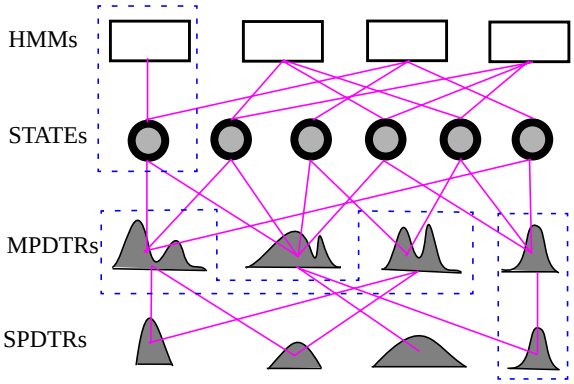

FIGURE 7 : Overall four-layer sharing structure of the speech recognition engine(SRE).

Notice that the items inside the dashed blocks may be shared. Also notice that the sharing can be performed across different layers: e.g., phone models and HMM states, MPDTRs and SPDTRs. Depending upon the recognition evaluation, this top-down or bottom-up procedure might continue until the best result is achieved.

3. 5000 US CITY NAMES EXPERIMENTS

TABLE 1 summarizes a set of experiments conducted to recognize isolated US city names. The speech data was recorded in a quiet sound booth at SONY Research Labs in San Jose, California. A Sennheiser HMD-410 headset microphone was used for all recordings. The vocabulary items for the training were selected to include the triphones and biphones from the 2000 most frequent words of English. This word list was composed by taking the most common words to all the following three published sources: Brown Corpus[5] top 5k words, the British National Corpus[6] top 5k words, and the Switchboard Corpus[7] top 5k words. To balance our training vocabulary, we also added a set of less common words from the above three sources and a set of randomly chosen words from a 50k dictionary.Finally, 75 city names and 25 car navigation commands were also added. The total number of unique recorded words for training only was 12,487.

The system was trained with speech data recorded from 140 speakers. The training data set has a total of 7164 triphones out of 75813 recorded tokens. The testing data contain 394 tokens from 4 testing speakers who were not included in the training data set. We use ~5000 US city names as the first set of experiments because it was easy to generate the phonetic dictionary with the city names database that exists at SONY. The actual recognition dictionary contains 4927 US city names and 25 commands for car navigation devices. Since the dictionary allows multiple pronunciations, the total number of phonetic transcriptions in the dictionary is 29,485.

The speech feature used throughout this paper was the conventional MFCC and its first and second derivative. A single 39 dimensional feature vector was used as well as the diagonal covariance matrix.

Results from experiments where the Gaussian function was used as the SPDTR (see FIGURE 7) are shown in TABLE 1 . The models in experiment 1 contain all the triphone models appeared in training. The models in experiment 1 have been clustered to a smaller number of models for experiments 2 and 3. The models in experiment 4 were generated from the models in experiment 3. Because of the state sharing, some extra models were generated for unseen triphones. This leads to 6865 unique models in experiment 4, which is higher than the 4122 models in experiment 3. The models in 5 were generated from experiment 4. The models in 6 were in turn generated from model 5; and the models in experiments 7, 8, and 9 were generated from experiment 6.

Experiment 2 as shown in TABLE 1 outperformed experiment 1 because the shared models provided more robust estimated parameters. Experiment 7 in turn outperformed experiment 5 with fewer Gaussians, showing the advantage of using the multi-layer sharing as depicted in FIGURE 7. Experiments 7 and 8 show the recognition performance with different numbers of states, MPDTRs, and SPDTRs. Even though there is a slight degradation in recognition performance from experiment 7 to experiment 8, we believe a proper balance of the number of MPDTRs should offer a better recognition result with fewer HMM parameters. Experiment 9 shows that a better HMM structure is achieved by sharing more MPDTRs and SPDTRs than in experiment 6.

#	lowest shared struct	#of model	#of state	#of MPDTR	#of gauss SPDTR	#of mix	top1 rec acc
1	NONE[a]	7213	10522	10522	10522	1	77.2%
2	model	1753	5259	5259	5259	1	83.8%
3	**model**	**4122**	**12366**	**12366**	**12366**	**1**	**80.5%**
4	state	6865	2975	2975	2975	1	77.7%
5	state	6865	2975	2975	5950	2	86.3%
6	**state**	**6865**	**2975**	**2975**	**11900**	**4**	**91.1%**
7	gauss	6865	2975	2975	5678	4	90.9%
8	gauss	6865	2177	2177	5186	4	90.1%
9	**gauss**	**6240**	**2975**	**2262**	**7841**	**4**	**91.4%**

a.) 7164 triphone models appeared in the training data set. The remaining 117,698 - 7164 unseen triphones were shared to monophones.

TABLE 1. Five thousand US city name recognition results with different HMM sharing structures.

The above results show the effectiveness of using the four-layer HMM sharing structure to improve the recognition while reducing the number of parameters. The time available to us limits our experiments to only continuous HMM recognition. However, we believe that the same HMM sharing structure will be equally effective on discrete HMM recognition experiments with the assistance of EQUATION 2.

4. CMU DICTIONARY EXPERIMENTS

Our focus, as stated in the title of this paper, is on very large isolated word recognition performance. A convenient English dictionary available for evaluation is CMU's 110k dictionary[4]. This dictionary contains mostly words with single phonetic spelling for each word with 39 basic phonemes.

Because isolated word recognition uses no grammartical or syntactical information, and no other high level knowledge except the low level acoustic information, it is very important that the speech dictionary describe the phonetic spellings as accurately and completely as possible.

A modified dictionary (D2), as compared to the CMU original dictionary (D1), allows optional closures before stops and optional glottal stop phones before vowels. Another version (D3), as compared to D2, corrected some of the phonetic spellings of the words and also added additional spelling variations based on a set of phonetic rules created on expert knowledge. The last improved version of the dictionary D4 comes from D3 plus the addition of real transcriptions from our training data set.

These four versions of 110k CMU dictionaries and the recognition accuracies with the best obtained HMMs are shown in TABLE 2 . The HMMs used in TABLE 2 are trained with 153 training speakers. Recognition is performed on 9633 word tokens from 20 independent speakers. These 9633 testing tokens contain 1783 unique common English words. For example, “coach”, “coast”, “coat”, “code”, “coke”, and “cold” are all included in the set for testing.

CMU dictionaries	D1 (original)	D2 +closure	D3 +spelling	D4 +training
top1	44.9%	66.4%	79.2%	83.9%
top2	53.5%	76.6%	87.4%	90.9%
top3	57.7%	80.7%	90.2%	93.3%

TABLE 2. Isolated word recognition results with 110K CMU dictionaries.

It can be seen from the table that we have significantly improved the large vocabulary isolated word recognition accuracy by improving the CMU dictionary. It is very interesting to see that D2 outperforms D1 by about 23% by simply introducing optional closure to the dictionary. The reason may be partially due to the importance for isolated word recognition to have a dictionary that accounts for a very detailed acoustic-phonetic variations. The better performance of D3 and D4 shows the significant impact of a good phonetic dictionary on recognition accuracy.

5. OTHER DICTIONARIES

Our goal is to construct a speech recognition engine general enough for any vocabulary. To check how well we have achieved this goal, we re-evaluated the newly-developed HMMs reported in section 4 using various dictionaries. The dictionaries used are the 5000 US city names dictionary(D7), the 5000 most frequent English words dictionary(D4) created at SONY, and the 50,0000 English dictionary(D5) developed at SONY. (Though the 50k phonetic dictionary was developed at SONY, the 50k word list originated from an English-Japanese bilingual dictionary[8].) For all these other experiments, the model remains unchanged. The data to test D3, D5, and D6 is the same as the data as reported in section 4. The data to test D7 consists of 6126 words of 16 independent speakers. These 6126 words contains 75 unique city names and 25 navigations commands. The top3 recognition performance is shown in TABLE 3 .

dictionaries	110k D3	50k D5	5k D6	5k cities D7
top1	79.2%	79.6%	85.9%	87.0%
top2	87.4%	88.6%	91.2%	91.1%
top3	90.2%	91.4%	92.3%	92.6%

TABLE 3. Isolated word recognition results with dictionaries of different sizes and different types.

TABLE 3 shows that the top1 recognition accuracy varies quite dramatically for different dictionaries, but the top3 recognition accuracies have less variation. Our future work is to further improve the top 3 recognition results for these different dictionaries.

6. CONCLUSION

We have developed a Speech Recognition Engine (SRE) with integrated discrete and continuous HMMs in such a way that they can effectively use the same four-layer parameter sharing structures. We have used the system to conduct experiments to improve very large vocabulary isolated word recognition. Experiment results show that we have been able to significantly improve 110k isolated word recognition accuracy while maintaining robustness for several different vocabularies of varying sizes.

7. REFERENCES

1. K.F. Lee, H.W. Hon (Nov. 1989). Speaker-Independent Phone Recognition Using Hidden Markov Models. *IEEE Trans. ASSP,* Vol. 37 (1641-1648)
2. S. J. Young and P.C. Woodland (Oct. 1994). State Clustering in Hidden Markov Model-Based Continuous Speech Recognition. *Computer Speech and Languag*e, 8 (369-383)
3. A. Ljolje (April 1994). High Accuracy Phone Recognition Using Context Clustering and Quasi-Triphonic Models. *Computer Speech and Language*, 8 (129-151)
4. R. Weide, et al(Nov 1995). CMU dictionary(v. 0.4) README. *http://www.speech.cs.cmu.edu/cgi-bin/cmudict*.
5. Kucera, et al(1967). Brown Corpus. *Computational analysis of present-day American English Providence*, Brown University Press.
6. Adam Kilgarriff, et al(Mar 1996). British National Corpus. *ftp://ftp.itri.bton.ac.uk/pub/bnc.*
7. Byrne Bill, et al(1996). WS96 Switchboard Data Resources (1996). *ftp://homer.clsp.jhu.edu/pub/swbdWS96.*
8. Yoshio Koine, et al(1980). Kenkyusha’s new English-Japanese dictionary, 5th ed., Tokyo, Japan.