IMPROVED HMM PHONE AND TRIPHONE MODELS FOR REALTIME ASR TELEPHONY APPLICATIONS

Ilija Zeljković[1] *and Shrikanth Narayanan*[2]

AT&T Laboratories
600 Mountain Avenue, Murray Hill, NJ 07974
[1]ilija@research.att.com [2]shri@research.att.com

ABSTRACT

Development of human-machine dialog applications for messaging and information retrieval over the telephone pose stringent requirements on accuracy and speed of the automatic speech recognition (ASR) system. In this paper, we describe strategies for improved acoustic-phone modeling directed toward increasing recognition accuracy while maintaining the number of phone units low. Specifically, this paper considers: (1) The development of an improved set of head-tail context-dependent (CD) triphones. (2) A novel criterion for better selection of the number of states assigned to each phone unit based on the coefficient of variation measure of feature components in HMM-Gaussians. Performance of the models is evaluated using data that represent real telephony applications.

1. INTRODUCTION

Acoustic phoneme-based models form the basis of many current HMM-based ASR systems. Improved phone-based recognition can be achieved by using CD models. Since the number of possible CD models is typically large (for example, for a 41 CI-phone set, the number of CD triphones = $41^3 = 68921$), recognition efficiency is compromised. However, improved efficiency can be obtained by modifying the triphone modeling. One approach is to model contextual effects only on the head and tail of the triphone model, keeping the body common for all contexts of the same phone. This approach has already proved successful in connected digit recognition [3] and is extended to subword phone modeling in [8] and in this paper. For a basic set of 40 phones and silence (which is context independent), the total number of preceding and following CD phones is 2x40x41 = 3280. Furthermore, in practice, not all contexts appear. Broad classification based on the frequency of occurrence of these phones in a generic training database resulted in null, high, and low frequency groups. For example, in a generic phonetically-rich database we considered, only 64% of the CD phones were present. Of those, 70% deemed highly frequent ($>$ 16 occurrences) and were considered for modeling. Low-frequency contexts were combined to build CI heads and tails. The use of task-specific databases will of course further refine the CD model building. Phone units were modeled using either 3 or 5 state left-to-right HMMs. Assignment of the 'optimal' number of states for each unit was based on a novel criterion that used the average coefficient of variation (ACV) of feature components in the HMM-Gaussians, resulting in significant improvement in accuracy.

In Section 2 the basic HMM structure used in the recognizer is described. Descriptions of the ACV criterion and the CI/CD phone units are given in Sections 3 and 4, respectively. The experimental results and conclusions are given in Section 5.

2. THE HIDDEN MARKOV MODEL BASED SPEECH RECOGNIZER

The speech units (words and subwords) as well as silence and impulse-type noise are modeled by first order, left-to-right HMMs [2]. The continuous probability density function for the state observation vector, $b_j(\mathbf{O})$ is represented as a mixture of a finite number of Gaussians of the form

$$b_j(\mathbf{O}) = \sum_{m=1}^{M} c_{mj}\mathbf{N}[\mathbf{O}, \mu_{mj}, U_{mj}] \qquad (1)$$

where c_{mj} is the mixture weight for the m^{th} component in state j, $\mathbf{N}$ represents a multivariant normal density, μ_{mj} is the mean vector for mixture m in state j, and U_{mj} is the diagonal covariance matrix for mixture m in state j. The observation vector consists of 39 features: 12 cepstral coefficients, dynamically normalized energy as well as their first and second derivatives. Features are generated from a linear prediction analysis. The HMM parameters c, μ, and U are estimated from labeled speech using a segmental k-means algorithm [4]. State boundaries in each training token are determined by optimal (Viterbi) alignment of the current model with the token. All frames from all training data assigned to a particular state are clustered into M clusters. From the vectors within each cluster, μ_{mj} and U_{mj} are calculated, and the mixture weight is derived from the number of frames in a cluster and the total number of frames for the state.

In the recognition process, the sequence of observation vectors of an unknown speech utterance is matched against a set of stored hidden Markov models representing phones. A search network is generated by a finite state grammar that describes the valid set of phone strings (utterances). The network search algorithm, described in [2,5], returns

the phone string candidate (sentence) that gives the highest probability.

3. AVERAGE COEFFICIENT OF VARIATION FOR HMMS

One way to assess the quality of the HMMs is to compare the average standard deviation of all its feature components over all Gaussian mixtures of all states. Since the absolute values of feature means, μ, vary greatly across feature categories (such as energy, cepstrum, delta-cepstrum), Gaussian mixtures, and HMM states, one would intuitively expect that a better measure of quality is the standard deviation of each component normalized by the absolute value of the corresponding mean. Such a measure has been successfully used to assess the performance characteristics of industrial products [1]. Coefficient of Variation (CV) of feature component f, a dimensionless quantity, is defined as $r_f = \frac{\sigma_f}{|\mu_f|+\epsilon}$ where σ_f is the standard deviation and μ_f the mean of the particular feature component f. ϵ is a small number ($\epsilon = 0.0001$ for our purposes) to avoid division by zero. It is more meaningful to average all Coefficients of Variations over all features in the particular HMM or in the whole set of HMMs yielding the Average Coefficient of Variation, $ACV = [\frac{\sigma}{|\mu|+\epsilon}]$.

NUMBER OF HMM-STATES PER PHONE: Phoneme units, in our recognizer, are typically modeled with 3-state left-to-right HMMs. Since phones represent a linguistical rather than a stationary acoustic segment, 3 states may not be enough to model phonemes with diverse acoustical segments. Thus, 5-state HMMs for phoneme units are examined, and the ACV value is used to automatically select a 3-state or 5-state representation for each unit.

NUMBER OF GAUSSIAN MIXTURES FOR EACH HMM STATE: Traditionally all HMM states are modeled with same number of Gaussian mixtures. However, the ACV values for background HMMs were found to be as much as twice larger than the ACV values for phone HMMs. For example, using data from a corpus comprising 17500 phonetically-balanced sentences of telephone speech, we found that the mean ACV values for the speech phone sets are fairly close (9.0 for 3-state, 32 mixture HMM and 8.8 for 5-state 19 mixture HMM[1]) while the ACV values for background (silence, noise) HMMs are dramatically higher: 23.1 for 1-state silence and 21.7 for 3-state noise. Each state in the background HMM is represented by 32 mixtures. In order to reduce the ACV values, the influence of using more Gaussian mixtures per state is considered. By increasing number of mixtures in background HMMs their ACV values decrease (Table 1) and the recognition performance improves. For example, in a connected digit recognition task, the improved HMMs using the ACV measure resulted in error reduction rates ranging from 32% to 48%, for different telephone databases.

Results on movie name review task:
The performance of the different phone and background HMMs is evaluated on 1111 utterances containing a movie name. The speech data was collected during a technical trial of movie review service. Each utterance contains an actual movie name spoken by an adult person. The whole set of allowed phrases consists of 168 movie names (representing 121 unique titles) which were showing in theaters at the time of the trial. English text for each movie name is transcribed into a string of phones using the AT&T Text-to-Speech front-end. The recognizer's output phone string is restricted by the grammar to one string representing movie names. The recognition results are given in Table 2.

[1]Note that the total number of Gaussians for the 3-state and 5-state models are about the same.

No. of Gaussians	ACV	
	1-state	3-state
32	23.1	21.7
64	19.6	14.8
128	16.7	12.9
256	13.3	11.8

Table 1: ACV values for silence HMMs.

No. of mixtures/state		Phrase error rate %		
1-state silence	3-state noise	3-state phones	5-state phones	3/5-state phones
32	32	5.9	4.4	4.1
64	64	5.3	4.1	3.6
128	128	5.2	3.9	3.4
256	256	4.5	3.6	3.0

Table 2: Phrase Error Rate for Movie Review Task.

The results in Table 2 show the clear improvement in recognition accuracy when the phone HMM structure change from 3 to 5 states per phone and also when the phone structures are combined in the way that the phone-set ACV value is decreased. The accuracy also significantly improves when much larger number of Gaussians is used in the background HMMs i.e. when the background HMM's ACV value decrease to the average phone ACV value.

4. PHONE UNITS

Context Independent Phones: Arriving at an optimal definition of a subword-unit set is a rather complex goal, depending both on training data and the required speech recognition task [6]. For small vocabulary telephone services the most generally used subwords are context independent *phonemes*. In our tests, forty phonemes are used to transcribe the training and test vocabularies. Transcription is performed word by word through dictionary lookup. During transcription the phonemic context is disregarded, allowing a one-to-one correspondence between phonemes and phones (the acoustical realizations of phonemes) to be retained. There are two reasons that justify usage of a 40-phone set, rather than a 46 phone set [7, 6]. First, most of the omitted phones are produced in inter-word contexts and most telephone services have single word or short sen-

tence input where these inter-word contexts are minimal. Second, keeping the number of CI phones small is crucial to reducing the number of CD phone models, as this is proportional to square or cube of the number of CI phones.

Context-Dependent Task-Independent Phones: A further improvement of phone based speech recognition can be achieved by using CD instead of CI phones [10]. With CD phones, a separate variation of each basic phone can exist for each combination of preceding and following phones [6, 7]. Unfortunately, the total number of possible CD phones is very large, equal to the third power of the number of basic CI phone units. In our case, a total of $41^3 = 68921$ CD phones (triphones) could exist. Although in practice all triphone combinations do not appear in English, the total number is still very high. For example, the total number of different triphones in our training database is 10514. This large number of phone units presents several practical problems. The collection of a sufficient amount of training data is one problem, and increased memory and processing time required for performing recognition is another. In telephone services, memory and processing time are especially valuable.

Instead of modeling each triphone as a fully separate model for each context, one can consider that only its beginning (*head*) and its end (*tail*) is affected by preceding and following context. That means that the triphone *body* can be shared for all contexts of the same phone. In this case the possible number of CD units is equal to two times the square of the number of CI units, and middle phone part (*body*) is common for all CI phones. In our work, a set of 40 base phones are used, most of them modeled with 5 HMM states, as compared with 46 base phones and 3 state HMMs that are typically used. It is possible to reduce further the number of context dependent units by grouping phones into broader phonemic classes. Instead of having a separate CD model for each phone in the group, one CD model for the whole group can be used.

With a basic set of 40 phones and *silence*, the total number of preceding and following contexts is 2 x 40 x 41 = 3280. (In this paper, *Silence* is treated as context independent.) In the English language some of the above contexts will never appear. In our training database (see Section 5), a total of 2053 different contexts appeared at least once. The 1402 head and tail contexts that appeared at least 16 times in the training phrases were modeled. Low frequency contexts were combined to build context-independent heads and tails.

In order to provide general task-independent phone units that can be used in any telephone service, phone-HMMs were trained on a generic phonetically-rich database without regard to any specific application. For established services, however, it is beneficial to retrain (or adapt) HMMs on a task-specific database as argued in [6].

It was pointed out in Section 3 that 5-state phone HMMs give higher accuracy than 3-state models and that combination of 5-state and 3-state phone models outperforms 5-state phone models. The ACV measure is used to automatically select a 3-state or 5-state representation for each unit. Two HMMs are used to represent the non-speech signal. One, a single-state HMM, is intended to represent stationary background conditions (silence). The other, a three-state HMM, is meant to model short acoustical noise, as well as coughing and breath noise. This HMM is referred to as a noise model.

The CD phone bodies are modeled with 1 state if the corresponding CI phone HMM has 3 states, or with 3 states if the corresponding CI phone has 5 states. All states in phone bodies are modeled with 19 Gaussian mixtures. However, there was not enough data to give satisfactory estimates for the mean and variance of all 19 Gaussians for these context-dependent heads and tails. On average, heads and tails are modeled with 13 mixtures. As mentioned in Section 3, the 1-state and 3-state silence HMMs are modeled with 128 Gaussians/state as a compromise between low ACV values and computational efficiency.

5. EXPERIMENTAL RESULTS

5.1. TRAINING PHONE HMMS

HMMs for phone units are trained on speech consisting of 12,146 short phrases recorded by approximately 300 speakers over long distance telephone lines from different dialectal regions in the United States. A list of generic phrases was created with a 'greedy' algorithm [9] that picked items according to high triphone entropy, so that successive phrases contained as many new triphones as possible. The phonetic transcription for each training utterance was performed automatically based on the orthographic (or spelling) transcription of the speech using the AT&T Text-to-Speech preprocessor.

5.2. TEST RESULTS

Town name recognition:
The performance of CD and CI phone HMMs is evaluated on 4846 utterances containing a town name. The list comprised 1221 town names in the state of New Jersey. The speech data was collected from 100 different speakers. and each speaker spoke 50 town names. The transcription from the orthography into a string of phones was done using the AT&T Text-to-Speech front-end. The recognition results are given in Table 3. In both model sets background is modeled by a single-state, 128-mixture HMM as well as a 3-state/128-mixture HMM. The results show a 25% error reduction when the context-dependent (head-body-tail) phone HMMs are used instead of context-independent HMMs trained on the same database.

Short word recognition:
Correct recognition of short words containing one or two syllables (for example, *yes, no, help*) is important for many interactive systems. Results comparing CI and CD models on a data set of 125 short words is given Table 3. The data were collected from an experimental trial with 25 talkers. Results show significantly better performance in the CD

Task	Error rate (%)	
	Context independent	Context dependent
Town names	13.9	10.5
Short words	16.8	6.4
Short strings	13.6	8.3
Long strings	5.6	4.2

Table 3: Word error rates (%) for various recognition tasks.

cases.

Short string recognition (1-8 words; for example, *get me the operator*):
This data consists of 1500 messaging command phrases (2 to 8 words long) spoken by 300 talkers over long distance telephone lines from various regions in the United States. The test results given in Table 3. Note that the search was restricted by a finite state grammar network.

Long string recognition (10-25 words; for example, *send this message to 1 1 1 2 2 2 3 3 3 3*):
In this case we consider strings that are a combination of English words and digits. Results using CD/CI subword models (without any special regard to the embedded digit strings) were somewhat worse than that of the short string recognition. However, when a hybrid combination of subword phone models and whole word digit models was used much better results were obtained (Table 3). Other testing conditions were similar to that used for the short string recognition.

6. CONCLUSIONS

Context dependent, task independent head-body-tail phone units, where only heads and tails are context dependent, were shown to out-perform corresponding CI phones. Results were particularly impressive for short words. Improved performance could also be obtained by using appropriate combination of whole word models with subword models. In further evaluation of CD phone performance, a larger training database will be used so that lower frequency contexts can be trained. To further reduce the number of CD units, the performance of CD units where similar contexts are combined will be examined.

In terms of total number of mixtures the size of CD HMMs is about 6 times bigger than the size of CI HMMs: 22619 compared to 3863 Gaussian mixtures.

7. REFERENCES

[1] Taguchi G., "Off-Line and On-Line Quality Control Systems," *Proc. Int. Conf. on Quality Control*, Tokyo, Japan, 1978.

[2] Rabiner, L. R., Wilpon, J. G. and Juang, B. H., "A Model-Based Connected-Digit Recognition System Using Either Hidden Markov Models or Templates," *Computer, Speech and Language*, Vol. 1, No 2, pp. 167-197, December 1986.

[3] Wilpon J. G. et al, "Connected Digit Recognition," *AT&T Bell Laboratories Technical Memorandum*, November 1993.

[4] Rabiner, L. R., and Wilpon, J. G., Juang, B. H., "A Segmental k-means Training Procedure for Connected Word Recognition Based on Whole Word Reference Patterns," *AT&T Tech. Journ.*, Vol. 65, No. 3, pp 21-31, May 1986.

[5] Lee C. H., Rabiner, L. R., and Juang, B. H., "Connected Word Recognition Using a Frame Synchronous Network Search Algorithm" *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 1649-1658, November 1989.

[6] Lee C.-H., Gauvain J.-L., Pieraccini R. and Rabiner L. R., "Large Vocabulary Speech Recognition," *Speech communication*, Vol. 13, pp. 263-279, 1993.

[7] Lee, C.-H. Giachin, E., Rabiner, L. R., Pieraccini, R., and Rosenberg, A. E. (1992) "Improved Acoustic Modeling for Large Vocabulary Continuous Speech Recognition," *Computer, Speech & Language*, Vol. 6(2), pp. 103-127.

[8] Grubbe, R., "Personal Communication."

[9] van Santen, J. P. H., "Perceptual experiments for diagnostic testing of text-to-speech systems," *Computer Speech and Language*, Vol 7, pp 49100, 1993.

[10] Katunobu I., Satoru H., Hozumi T., "Continuous Speech Recognition by Context-dependent Phonetic HMM and an Efficient Algorithm for Finding N-best Sentence Hypotheses," *Proc. ICASSP 92*, Vol. 1, pp. 21-24, March 1992.

[11] Roth R. et al, "Large Vocabulary Continuous Speech Recognition of Wall Street Journal Data," *Proc. ICASSP 93*, Vol. 2, pp 640-643, April 1993.