6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

OPTIMIZATION OF UNITS FOR CONTINUOUS-DIGIT RECOGNITION TASK

Sachin S. Kajarakar[1] **and Hynek Hermansky**[*,**]

[*]**Oregon Graduate Institute of Science and Technology**, Portland, Oregon, USA.
[**]**International Computer Science Institute**, Berkeley, California, USA.

ABSTRACT

The choice of units, sub-word or whole-word, is generally based on the size of the vocabulary and the amount of training data. In this work, we have introduced new constraints on the units: 1) they should contain sufficient statistics of the features and 2) they should contain sufficient statistics of the vocabulary. This led to minimization of two cost functions, first based on the confusion between the features and the units and the second based on the confusion between the units and the words. We minimized first cost function by forming broad phone classes that were less confusing among themselves than the phones. The second cost function was minimized by coding the word-specific phone sequences. On the continuous digit recognition task, the broad classes performed worse than the phones. The word-specific phone sequences however significantly improved the performance over both the phones and the whole-word units. In this paper we discuss the new constraints, our specific implementation of the cost functions, and the corresponding recognition performance.

1. INTRODUCTION

Speech recognition systems map sequence of feature vectors to the corresponding sequence of words. Here feature vector refers to a set of measurements from speech signal, e.g., logarithmic spectrum. These feature vectors are first converted to an intermediate representation which is referred as a set of "units" in this paper. A particular sequence of one or more units forms a word. A "dictionary" represents all the pronunciation of the word in terms of these units. Different speech recognition systems use different units to represent the words in its vocabulary. For example, small vocabulary (digit recognition) systems use a set of whole-word units, and medium and large vocabulary systems use a set of sub-word units. These sub-word units are derived from the basic phonemes of the language [1]. In general, the choice of units is based on the size of the vocabulary and the amount of training data [2]. Although this choice has been adequate, we believe that a set of optimal units must depend on the confusability of the words in the vocabulary in addition to the size of the vocabulary.

The problem of automatic derivation of sub-word units was addressed by numerous researchers [3, 4, 5, 1, 6]. However most of this work was aimed at the medium or the large vocabulary tasks. For these tasks the sub-word units were derived from the tree-based clustering [5] of the context-dependent phones[1] (CDPs). This technique however is based on acoustic similarity between the CDPs and ignores the inherent confusability of the words in the vocabulary.

Note that the aim is to improve the robustness of the speech recognition system and robust feature selection [7, 8] is a possible approach too. However we have pursued an alternative approach in this work. Given a set of features, we have derived a set of new units for the given (continuous digit recognition) task. The new units were derived using an idea that the optimal units must contain sufficient statistics of the vocabulary and the features. This concept is explained in section 2 and it leads to two cost functions. We describe the experimental setup in section 3. This is followed by the implementation of the cost functions and the corresponding recognition results in section 4 and 5 respectively. We conclude the paper with discussion of results in section 6.

2. PROBLEM FORMULATION

Let $\mathbf{F} = \{f_i\}$, where $i = 1, .., F$, represent the sequence of feature vectors derived from the speech signal. Let $\mathbf{U} = \{u_i\}$, where $i = 1, .., U$, be the corresponding set of sub-word units and $\mathbf{W} = \{w_i\}$, where $i = 1, .., W$, be the corresponding set of words. The optimal set of units, $\tilde{\mathbf{U}}$, must be such that $\mathbf{F}$ and $\mathbf{W}$ must be conditionally independent given $\tilde{\mathbf{U}}$, i.e., $p(f, w|\tilde{u}) = p(f|\tilde{u})p(w|\tilde{u})$.

We explain this requirement in the following discussion. Consider the training phase of the speech recognition system when the words are represented by the units and the units are modeled using the features. This is represented by the sequence $\mathbf{W} \longrightarrow \mathbf{U} \longrightarrow \mathbf{F}$. In this case, the relation between the information ($I(\mathbf{W};\mathbf{F})$) transfered from the words to the features and the information ($I(\mathbf{W};\mathbf{U})$) transfered from the words to the units is given by

$$I(\mathbf{W};\mathbf{U}) \geq I(\mathbf{W};\mathbf{F}).$$

Thus $I(\mathbf{W};\mathbf{F})$ is bounded by $I(\mathbf{W};\mathbf{U})$ [9]. The only way to survive this bottleneck is by assuming equality in the above equation. This condition is achieved if $\mathbf{U}$ contains sufficient statistics for $\mathbf{F}$ or vice versa, i.e.,

$$H(\mathbf{F}|\mathbf{U}) = H(\mathbf{U}|\mathbf{F}) = 0, \tag{1}$$

where $H(\mathbf{F}|\mathbf{U})$ or $H(\mathbf{U}|\mathbf{F})$ denote the conditional entropy. $H(\mathbf{F}|\mathbf{U})$ represents the uncertainty about $\mathbf{F}$ given that $\mathbf{U}$ is known.

Now consider the testing phase, when the features are converted to the words by recognizing the corresponding sequences of units. The operation is described by the sequence $\mathbf{F} \longrightarrow \mathbf{U} \longrightarrow \mathbf{W}$. Using the a similar argument that we

[1] "phone" refers to context independent monophone

made above, we get another condition for the optimal units,

$$H(\mathbf{W}|\mathbf{U}) = H(\mathbf{U}|\mathbf{W}) = 0, \qquad (2)$$

i.e., the units should contain sufficient statistics of the vocabulary. Combining these two conditions (equation (1) and (2)), we can say that the optimal units must contain sufficient statistics of the features as well as the vocabulary.

In this work, we used both the conditional entropies (from eqn.(1) and(2)) separately as the cost functions for obtaining new units. The general procedure was 1) start with phones and estimate $H()$; 2) combine two or more phones and reestimate $H()$; 3) find the combination of phones that minimizes $H()$ and replace the phones with the new unit; and 4) repeat this procedure till it meets the stopping criterion.

3. EXPERIMENTAL SETUP

Before discussing the solutions, we will briefly describe the experimental setup in this section.

We have used the digits part of OGI Numbers [10] database for the experiments. This part contains different recordings of continuous digits. These recordings are in the form of zip-code, street address or other numerical information over different telephone numbers. Approximately 2547 files were used for training and 2167 files were used for testing. We used energies from 15 bark-warped filter-banks compressed using log non-linearity as features. The features were estimated from 25 ms of speech waveform , using hamming window, every 10 ms. These features were normalized by removing mean over each filter-bank trajectory and were further appended by the time derivative (Δ) and double derivative ($\Delta\Delta$) computed using 50 ms and 90 ms time window respectively. Finally the resulting 45 dimensional feature vectors were whitened using the KL transform which was estimated from the training data.

The baseline systems contain 23 phone units (dictionary in Table 1) and 12 whole-word units. The phones were modeled using 5 state, 3 component mixture HMMs and whole-word units were modeled using 16 state, 3 component mixture HMMs. When sequence of 2 phones was labeled as one unit, it was modeled using 10 state, 3 component mixture HMM. Similarly sequence of 3 phones was modeled using 15 state, 3 component mixture HMM when the sequence was represented as one unit.

4. SOLUTIONS

Equations (1) and (2) indicate that we can optimize the phone units by considering either the features or the words and it should lead to the optimal solution. There can be many ways of optimizing the units. We have only presented one possible scheme in this paper.

4.1. Using $H(\mathbf{F}|\mathbf{U})$

$H()$ from equation (1) represents the uncertainty in the features given the units. We calculated $H()$ from the phone confusion matrix as follows. First 5 state, 3 component mixture HMM was estimated for each phone using the labeled training data. Using these HMMs, a phone recognition experiment was then performed on the training data to get a new set of phone labels. The new phone labels were compared to the original phone labels and a phone confusion matrix was formed. This confusion matrix was converted to the joint probability distribution function (PDF) by dividing it with the total number of phone segments. Finally $H()$ was calculated from this joint PDF ($H(\mathbf{F}|\mathbf{U}) = H(\mathbf{F},\mathbf{U}) - H(\mathbf{U})$).

$H()$ was reduced by finding new units which are less confusing among each other then the phones. In this work, we obtained new set of units by merging a pair of phones that resulted in the highest reduction in $H()$. For example, "ao" and "ow" were identified as the most confusing phones in the first step and were merged to form a new class ao_ow. The merging was repeated till all the phones were paired. The sequence of resulting merges and the corresponding $H()$ is shown in Table 2. Note that none of these merges resulted in identical pronunciation for two words.

After analyzing the sequence using the dictionary (Table 1), we observed that the merges are not independent of the dictionary. For example, the first merge (ao and ow) was influenced by the fact that two pronunciations of "four" differed by "ao" and "ow". Similarly most of the different pronunciations of seven are due to "ah" replaced by "eh". Consequently "ah" and "eh" were merged into a single unit in the following steps. Acoustic similarity between different phones was also influencing the merges, e.g., merging of "kcl" and "tcl" into one unit, merging of "s" and "th" into one unit, etc.

The recognition performance obtained using these units is also shown in Table 2. Although there was no significant[2] reduction in the performance till step 5, it was observed that reducing $H()$ from equation (1) did not improve the recognition performance. We will discuss this result further in the last section.

4.2. Using $H(\mathbf{U}|\mathbf{W})$

$H()$ from equation (2) represents the uncertainty about units given words and

$$H(\mathbf{U}|\mathbf{W}) = \sum_{w_i \in \mathbf{W}} p(w_i) * H(\mathbf{U}|w_i).$$

Therefore, minimizing $H()$ is equivalent to minimizing $H(\mathbf{U}|w_i)$ for each w_i assuming $p(w_i)$ is constant. In other words minimizing $H()$ is same as coding word-specific phone sequences.

For this minimization, we considered multiple pronunciations of a single word and estimated phone sequences that reduced $H(\mathbf{U}|w_i)$. These sequences were labeled as the new units. For example there are two pronunciations of "one" - "w ah n" and "w ah n ah" (see Table 1) and $H(\mathbf{U}|one) = 1.55$[3]. But "w ah n" is common in both the pronunciations. We created a single unit called "w_ah_n" from this sequence. The corresponding $H(\mathbf{U}|one)$ reduced to 0.918. The recognition error using these units also reduced to 5.6% (column 3, Table 3). The procedure was repeated for all digits and the modified dictionary is shown in Table 4. Note that we would get $H(\mathbf{U}|one) = 1$ by modeling the pronunciations as different units.

The new units were different for different digits. Only one new unit - "ey_tcl" - was formed for "eight" and 3 new units were formed for "seven". "Five" had only one pronunciation so it was modeled as a whole-word unit. The two

[2] significance is measured at $\alpha = 2.5\%$ throughout this paper

[3] we have used "bits" to represent $H()$ in this paper

pronunciations of "six" differed by insertion of "kcl" and the common sequences - "s ih" and "k s" - were modeled as single units. It is interesting to note that in the multiple pronunciations of digits, a vowel or a diphthong is replaced by another vowel or diphthong most often whereas the consonants are relatively stable in their position.

Continuous-digit recognition performance of the new units is compared to both the phones and the whole-word units in Table 4. The whole-word units performed significantly better then 23 phones and the new units outperformed phone units and whole-word units. These results are further discussed in the next section.

5. DISCUSSION

In this paper we introduced the concept that the units for speech recognition must contain sufficient statistics of the vocabulary and the features. This led to the definition of two cost functions and we aimed to obtain the optimal set of units by minimizing either one of them.

We interpreted $H(\mathbf{F}|\mathbf{U})$ (equation(1)) as a measure of confusion among the statistical models of the phones (section 4.1). We minimized it by merging the most confusing phones into one unit. The merges clearly lead to broad phone classes, viz., vowels, fricatives, stops, and silence. This is similar to the approach used by Shipman et. al.[11]. Shipman hypothesized that the preliminary acoustic analyzer can make a six-way distinction, viz., 1) vowels and syllabic consonants, 2) stops, 3) nasals, 4) strong fricatives, 5) weak fricatives and 6) glides and semivowels. On isolated-word recognition task, it was concluded that "one approach to isolated word recognition for a large lexicon may be to initially classify the sound units into several broad categories where the error in labeling is still small". In our case the formation of broader classes did not yield an improvement in the performance. This could be due to the fact that by merging two units, we were merging the phonetic context (e.g., "ey" and "iy") that was important for this task. We could have also blurred the transition between the words by merging the phones and thus reducing the continuous word recognition performance, e.g., merging "s" and "th" into one unit may erase the boundary between "six" and "three". In conclusion more work needs to be done to understand the effect of these phone merges on the decoding of the phone sequence.

The second cost function, $H(\mathbf{U}|\mathbf{W})$, represented the confusion between the units and the words. It was reduced by forming new units by combining word specific phone sequences. Note that $H(\mathbf{U}|\mathbf{W}) = 0$ for the whole-word models. Consequently they performed better than the phone units. They however fail to represent the multiple pronunciations of a word specifically the insertions and deletions at the phone level. The word specific phone sequences represented these effects and they further improved the continuous word recognition performance.

Finally we would like to investigate into the joint optimization of units based on the features and the vocabulary using appropriate constraints in the future work.

6. ACKNOWLEDGMENTS

We thank Prof. Fraser from Portland State University (PSU) for the discussions on this topic. The research was supported by NSA/DoD under MDA904-00-C-2089 ,NSF under IRI-9712579 and by an industrial grant from Intel Corporation.

7. REFERENCES

[1] C. H. Lee, J. L. Gauvain, R. Pieraccini and L. R. Rabiner, "Large Vocabulary Speech Recognition Using Subword Units ," in *Speech Communication*, 1993, pp. 263–279.

[2] K. K. Paliwal, "Lexicon-Building methods for an acoustic sub-word based speech recognizer," in *Proc. of ICASSP*, 1990, pp. 729–732.

[3] Li Deng and Don Sun, "A statistical approach to automatic speech recognition using the atomic speech units construucted from overlapping articulatory features," in *JASA*, 1994, pp. 2702–2719.

[4] George Saon and Mukund Padmanabhan, "Data-Driven Approach to Designing Compound Words for Continuous Speech Recognition," in *ASRU*, 1999.

[5] H. J. Nock, M. J. F. Gales, and S. J. Young, "Comparative Study of Methods for Phonetic Decision-Tree State Clustering ," in *Eurospeech*, 1993, pp. 111–114.

[6] R. Singh, B Raj and R. M. Stern, "Automatic generation of phone sets and lexical transcriptions ," in *ICASSP*, 2000.

[7] N. Morgan, "Temporal Signal Processing for ASR," in *ASRU'99*. IEEE, 1999.

[8] M. Hunt, "Spectral Signal Processing for ASR," in *ASRU'99*. IEEE, 1999.

[9] Thomas M. Cover and Joy A. Thomas, *Elements of Information theory*, John Wiley & Sons, Inc., 1991.

[10] R. A. Cole, M. Noel, T. Lander and T. Durham , "New telephone speech corpora at cslu," in *Proc. of EUROSPEECH*, Madrid, Spain, 1995, pp. 1159–1162.

[11] David Shipman and Victor Zue, "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems," in *Proc. of ICASSP*, 1982, pp. 546–549.

digit	pronunciation
one	w ah n
one	w ah n ah
two	t uw
two	tcl t uw
three	th ih r iy
three	tcl th r iy
three	th r iy
four	f ao r
four	f ow r
five	f ay v
six	s ih kcl k s
six	s ih k s
seven	s eh v ah n ah
seven	s ah v ah n
seven	s eh v ax n
seven	s eh v n
seven	s eh v ih n
seven	s eh v eh n
seven	s eh v ah n
eight	ih ey tcl t ah
eight	ey tcl t
eight	ey t
eight	ey tcl
nine	n ay n ah
nine	n ay n
oh	ow
zero	z ih r ah
zero	z iy r ow
zero	z ih r ow
sil	si

Table 1: Dictionary

Step	Merges	$H(\mathbf{F}\|\mathbf{U})$	% error
0	No merges (23 phones)	1.0892	6.2
1	ao + ow ⟶ ao_ow	1.0338	6.1
2	v + n ⟶ v_n	0.9921	6.5
3	eh + ah ⟶ eh_ah	0.9409	6.8
4	s + th ⟶ s_th	0.8947	6.9
5	tcl + kcl ⟶ tcl_kcl	0.8185	6.9
6	uw + ih ⟶ uw_ih	0.7976	7.7
7	ey + iy ⟶ ey_iy	0.7043	8.0
8	r + ay ⟶ r_ay	0.6747	9.0
9	k + t ⟶ k_t	0.6738	9.2
10	ax + w ⟶ ax_w	0.6382	9.3
11	z + f ⟶ z_f	0.6137	10.1

Table 2: Sequence of merges minimizing $H(\mathbf{F}|\mathbf{U})$ and the corresponding recognition performance

digit	NEW pronunciation	%error using new units
one	w_ah_n	5.6
one	w_ah_n ah	
two	t_uw	6.1
two	tcl t_uw	
three	th ih r_iy	5.8
three	tcl th r_iy	
three	th r_iy	
four	f_ao r	5.6
four	f_ow r	
five	f_ay_v	5.8
six	s_ih kcl k_s	5.9
six	s_ih k_s	
seven	s_eh_v ah_n ah	5.5
seven	s_ah_v ah_n	
seven	s_eh_v ax n	
seven	s_eh_v n	
seven	s_eh_v ih n	
seven	s_eh_v eh n	
seven	s_eh_v ah_n	
eight	ih ey_tcl t ah	6.0
eight	ey_tcl t	
eight	ey t	
eight	ey_tcl	
nine	n_ay_n ah	5.5
nine	n_ay_n	
oh	ow	6.2
zero	z_ih r_ah	5.9
zero	z_iy r_ow	
zero	z_ih r_ow	

Table 3: New units and the corresponding dictionary of digits obtained using $H(\mathbf{U}|\mathbf{W})$

units	phone	whole-word	from $H(\mathbf{U}\|\mathbf{W})$
% error	6.2	5.2	4.4

Table 4: Comparison of continuous-digit recognition performance using different units