From Phone Identification to Phone Clustering using Mutual Information [†]

P. O'Boyle, J. Ming, M. Owens, & F. J. Smith

School of Electrical Engineering & Computer Science
The Queen's University of Belfast, Belfast, BT7 1NN, Northern Ireland

E-mail: P.OBoyle@qub.ac.uk, J.Ming@qub.ac.uk, M.Owens@qub.ac.uk, FJ.Smith@qub.ac.uk
Tel: (+44 1232) 274536 Fax: (+ 44 1232) 666520

ABSTRACT

In this paper we show how a confusion matrix derived from phone identification experiments can be used to automatically generate phone clusters. These clusters can be applied when constructing triphone models to overcome the sparse data problem. Two techniques are presented; firstly an hierarchical clustering technique is described; then an open clustering technique is presented. Both of these use mutual information calculated on a probability distribution derived from the confusion matrix as a measure of phone similarity. Sample results from each technique are presented.

1. INTRODUCTION

Triphone based hidden Markov models are currently the most successful acoustic modeling approach in large vocabulary continuous speech recognition. However, due to the large number of distinct triphones and the limited availability of training data, clustering of training data is needed to produce effective models. The clustering can be achieved by combining training examples for triphones with *similar* context. Both subjective classifications[1] and broad phonetic classifications[2] have been employed for this purpose.

In this paper we propose two new methods of clustering triphone training data, based on an objective measure of phone similarity, derived from a confusion matrix. The measure is defined in terms of mutual information. Previous researchers have used data from a confusion matrix to examine errors and improve the performance of their models[3]. Here we demonstrate two objective approaches that utilize this data to form clusters.

Mutual information, calculated on a probability distribution derived from the confusion matrix, is used as a measure of phone similarity; that is, we calculate the change in mutual information, when the probabilities for distinct phones are combined, and use this as a measure of the similarity of the phones. We assume that the distinction between two phones is directly related to the decrease in mutual information which occurs when they are combined. Therefore a smaller decrease in mutual information indicates a smaller distinction and hence a greater similarity between the phones.

Two techniques are presented. The first constructs a fixed hierarchical classification using a greedy algorithm to approximate a classification which maintains maximum mutual information. The second uses an ordered list of similar phones to each phone, to allow data dependent clustering; in this way it can produce clusters with any required minimum number of training examples, and so ensures that accurate models can be trained.

2. CLUSTERING PHONES USING MUTUAL INFORMATION

2.1 The Confusion Matrix

A confusion matrix records a summary of the results from phone identification experiments. The confusion matrix records the counts $c(i, j)$ of the number of times a phone labeled as i is identified as phone j. For convenience we also define T as the total number of phone identification experiments, which is given by:

$$T = \sum_{i,j=1}^{n} c(i,j) \qquad (1)$$

where n is the number of distinct phones.

In our experiments the confusion matrix is produced using 61 phone models trained on the standard TIMIT

[†] This research was supported by EPSRC grant GR/K82505.

training set (1990 CD-ROM). The same training set is also used as test data for the phone identification experiments, to ensure that the confusion matrix produced is independent of the test set. A similar matrix can be produced from the test set; however, use of clusters derived from such a matrix could invalidate subsequent results, since the division between training and test sets would not have been maintained.

2.2 Mutual Information

The confusion matrix can be used to define the probability distribution p_{ij} that a phone is labeled as i and identified as j:

$$p_{ij} = \frac{c(i,j)}{T} \tag{2}$$

From this probability distribution the mutual information between the labeled and identified phones can be calculated. Let p_j^I be the probability that a phone is identified as j:

$$p_j^I = \sum_{i=1}^{n} p_{ij} \tag{3}$$

And let p_i^L be the probability that a phone is labeled as i:

$$p_i^L = \sum_{j=1}^{n} p_{ij} \tag{4}$$

Then the mutual information for the confusion matrix can be calculated as:

$$\begin{aligned} I(p^I; p^L) = & \sum_{i,j=1}^{n} p_{ij} \log_2(p_{ij}) \\ & - \sum_{j=1}^{n} p_j^I \log_2(p_j^I) \\ & - \sum_{i=1}^{n} p_i^L \log_2(p_i^L) \end{aligned} \tag{5}$$

The change in mutual information when data in the confusion matrix is combined can be used as a measure of the similarity between phone models (or phones). We calculate the change in mutual information when the recognition results for two models are combined. For example, if the results for phones j_1 and j_2 are combined, then the change in mutual information can be calculated as:

$$\begin{aligned} d(j_1, j_2) = & (p_{j_1}^I + p_{j_2}^I) \log_2(p_{j_1}^I + p_{j_2}^I) \\ & - \sum_{i=1}^{n} (p_{ij_1} + p_{ij_2}) \log_2(p_{ij_1} + p_{ij_2}) \\ & - p_{j_1}^I \log_2(p_{j_1}^I) + \sum_{i=1}^{n} p_{ij_1} \log_2(p_{ij_1}) \\ & - p_{j_2}^I \log_2(p_{j_2}^I) + \sum_{i=1}^{n} p_{ij_2} \log_2(p_{ij_2}) \end{aligned} \tag{6}$$

This corresponds to combining the results for phones *identified* as j_1 and j_2. A similar formula can be used to calculate the change in mutual information when results for phones which are *labeled* as i_1 and i_2 are combined. Here we use (6) as a measure of similarity when clustering phones, as this measures the similarity of the phone models. Similar results are achieved when the measure is based on combinations of phone labels.

2.3 Hierarchical Classification Using A Greedy Algorithm

A useful clustering scheme is one that maintains as much mutual information as possible. We could attempt an exhaustive search of all possible classifications and identify the best. However, even with a relatively small number of phones the number of possible classifications is extremely large.

We therefore use a greedy algorithm[4] to produce an approximately optimal classification. Thus the two phones which lead to the smallest reduction in mutual information are combined first. Then the pair with the next smallest reduction is identified, considering any previously combined pair as a possible candidate for combination. When a pair of phones is combined their probability components are simply added to produce an updated confusion matrix probability distribution. We continue combining phones until the required number of phone clusters have been produced.

Other methods can also be used to cluster the phones based on the same measure; for example, a top down approach[5] where clusters of phones are divided rather than combined.

2.4 Ordered Lists of Phones

When a fixed classification is used to combine training data with similar contexts, there is no guarantee that sufficient training examples will be available to train a reliable model. For example, using only 5 distinct classes, obtained from the above method, to cluster left context dependent phone models (for the TIMIT training set) 38 of the 189 models are assigned fewer than 10 training examples. In addition the fixed classification can produce a single model where multiple models could have been trained. We therefore propose a second clustering technique which overcomes these disadvantages.

We use the change in mutual information to produce an ordered list of phones similar to each phone. Training data is then merged based on these lists and subject to a minimum number of training examples needed to produce a model. Thus, any training data for a context dependent unit with insufficient training examples, to permit a single model to be built, is merged with other training data, from the most similar phone. We start with the least frequent contexts and continue merging until all clusters have attained the required minimum number of training examples. This data-dependent clustering permits a large number of models to be constructed while ensuring that each can be adequately trained.

Currently we employ this ordered list approach to classification to produce left and right context dependent phone models that are then combined to produce triphone models[6].

3. RESULTS

3.1 Hierarchical Classification

Table I shows a classification with 15 phone classes derived from a confusion matrix with 40 phone labels (the initial 62 phones having been folded to the standard 39 phones +q). The brackets indicate the order in which the phones and sub-classes were combined, e.g. n and ng were combined before being combined with m.

The classification in Table I seems reasonable, for example the nasals (m, n, and ng) have been combined into a single class, and the fricatives and africatives are placed into four classes. The classes also contain some less obvious pairings, e.g. [q, hh]. These may distinguish the automatically derived classes sufficiently from hand coded classes to permit improved performance in models based on the new classifications.

Table I

Automatically derived hierarchical classification with 15 classes

[b,p]	[f,th]	[y,iy]
[[d,t],[g,k]]	[s,z]	[aa,[aw,ow]]
[[dx,v],dh]	[m,[n,ng]]	[[[ae,eh],[ay,oy]],[ah,uh]]
[q,hh]	[l,w]	[ey,[ih,uw]]
[[jh,ch],sh]	[r,er]	sil

3.2 Ordered List Classification

Table II shows an example of the ordered list of phones for the phone *ch*. From this we can see the most similar phone to *ch* is *jh*; therefore data for a phone with context *ch* will be combined with data for the same phone with context *jh,* if necessary to reach the required threshold.

Table II

An example of the ordered list of phones for the context phone *ch*.

Context phone	**Ordered list of other phones**
ch	jh sh th t d uh z dx g dh p hh oy f s v q y k b ow aw ng eh ae w ay r ey m l aa uw iy er ah n ih sil

This and other similar lists are used to cluster the training data for each central phone in left and right context dependent phone models. An example of the clusters produced for right context dependent units of the phone *ah* is shown in table III. The available training examples are listed in the first major column and the derived classification is given in the second major column. From table III we can see that the single example of *ah* with context *ch* has been combined with other data for *ah* with contexts *z*, *sh*, and *jh*.

Using this ordered list clustering technique with a threshold of 100 training examples to produce clusters for left context dependent phones we produce 421 distinct clusters for the TIMIT training set. This compares favorably with the 189 clusters produced by a fixed classification scheme as described above in which 38 clusters had fewer than 10 training examples.

4. CONCLUSIONS

In this paper we have presented two new methods of clustering phones, both of which are based on measuring

the changes in mutual information derived from a confusion matrix. The hierarchical classification can be used in place of a hand coded classification and shows similar groupings. The technique can be used to produce any desired number of phone clusters, giving greater flexibility than a single fixed classification. The ordered list approach can be used to form context clusters dynamically specific to each phone. The number of clusters can be varied by selecting a threshold number of training examples that each cluster must contain.

This second technique is being applied to construct clusters of left and right context dependent phone units that are then combined to form triphone models[6]. This permits a less invasive use of context clusters as they are introduced only when there is insufficient training data to produce a model without context clustering, therefore retaining greater model resolution.

The suitability of these new clustering techniques is based on the assumption that phones that are easily confused will produce similar context effects. While this seems plausible, further research into this issue may be appropriate.

REFERENCES

[1] Ljolje, A. high accuracy phone recognition using context clustering and quasi-triphonic models., Computer Speech and Language, Vol. 8, 129-151, 1994.

[2] Deng, L., Lennig, M. Seitz, F. and Mermelstein, P. "Large vocabulary word recognition using context-dependent allophonic hidden Markov models", Computer Speech and Language, vol. 4., pp. 345-357, 1990.

[3] Deroualt, A-M, Context dependent phonetic Markov models for large vocabulary speech recognition., IEEE International Conference on Acoustics Speech and Signal Processing, 360-363, 1987.

[4] Brown, P. F., Della Pietra, V. J., deSouza, P., Lai, J. C., and Mercer, R. C. Class based n-gram models of natural language., Computational Linguistics, Vol. 18, 467-479, 1992

[5] McMahon, J., and Smith F.J., Improving statistical language model performance with automatically generated word hierarchies., Computational Linguistics, Vol. 22, 217-248, 1996

[6] Ming, J., O'Boyle, P., and Smith, F.J. "A Bayesian approach for building triphone models for continuous speech recognition", submitted to IEEE Transactions on Speech and Audio Processing.

Table III

Context Clustering for the Right Context Dependent Units of Phone *ah* Using the Ordered List Clustering Technique with a Minimum-Sample Threshold of 150.

(Phn=phone Ctx=context Freq=frequency)

Initial units						**Context clustered units**		
Phn	Ctx	Freq	Phn	Ctx	Freq	Phn	Ctx class	Freq
ah	sil	1751	ah	hh	81	ah	sil	1751
ah	n	842	ah	r	77	ah	n ng	933
ah	m	619	ah	sh	56	ah	m	619
ah	l	568	ah	y	8	ah	l aw ay ow aa	576
ah	s	530	ah	ih	5	ah	s	530
ah	v	433	ah	aa	3	ah	v	433
ah	z	248	ah	aw	2	ah	f th	309
ah	f	204	ah	jh	2	ah	z sh jh ch	307
ah	w	166	ah	ow	2	ah	hh y ih ey r er q	265
ah	dh	155	ah	er	2	ah	w	166
ah	dx	150	ah	ay	1	ah	dh	155
ah	th	105	ah	ey	1	ah	dx	150
ah	ng	91	ah	ch	1			
ah	q	91						