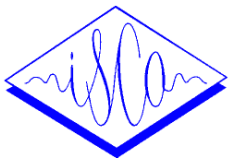

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

CLUSTERING TRIPHONES BY PHONOLOGICAL MAPPING

Maurice K. Wong

Verbex Voice Systems, Inc.
119 Russell Street
Littleton, MA 01460
USA

ABSTRACT

One of the most important issues in large-vocabulary continuous speech recognition is the modeling of subword units. To model context-dependent acoustic-phonetic variations, typically a large number of units such as triphones are used. Given a finite amount of training data, many triphones are underrepresented and remain undertrained or even untrained. This paper proposes an algorithm for mapping underrepresented triphones to adequately represented ones that are phonetically similar. First, all triphones are categorized according to their places and manner of articulation. Each triphone that needs to be mapped is compared to other triphones, and the candidates are ranked according to whether the left contexts and/or the right contexts are in the same phonetic class, as determined by acoustic-phonetic variations due to context. Second, if a good candidate has not been found, each candidate triphone is analyzed as phonological feature vectors, and the ranking of similarity is determined by the dot product of the vectors. The best candidate for mapping is chosen on the basis of phonetic similarity as well as the frequency of occurrence of the candidate triphones. In a recognition test of the Resource Management task, using this phonological mapping reduces the word error rate significantly.

1. INTRODUCTION

In order to model context-dependent acoustic-phonetic variations, large vocabulary continuous speech recognition systems typically make use of a large number of subword units such as triphones. Although this approach has helped to achieve high recognition accuracy, it does have a number of disadvantages. First, a large number of triphones is required for any large vocabulary. On the one hand, having models that are more specific means more sensitivity to contextual variants and thus higher recognition accuracy; on the other hand, the more specific the models are, the less trainable the system is, since the amount of training data is finite. There will always be triphones, particularly interword triphones, that are underrepresented in any corpus, and they tend to be undertrained, or in the worse case, untrained. Because the parameters of hidden Markov models (HMMs) can be robustly estimated only if training data are sufficient, gaps in the corpus can be detrimental to the outcome. A second problem involved in the use of triphones is the incorrect assumption that all triphone contexts are different, when in fact many of the contexts are acoustic-phonetically similar. This results not only in redundant triphone models, but also in separate models that are less than well-trained, when they should be merged into a single model that is more robust.

The most common solution for the problem with triphones that are underrepresented in the corpus is to revert to the corresponding diphone or context-independent phone. This is a data-driven approach and has the disadvantage of not being able to capture the context-dependent variations originally intended by the use of triphones. The context-independent phone or diphone employed as a substitute is less sensitive to specific contexts and may well be the cause of a recognition error. A different solution is to find a triphone that is similar to the one in question, based on its phonetic characteristics, such as using a decision tree to model an unseen subword unit [1]. The latter approach may be viewed as a theory-driven one, since a top-down decision is made on the choice of the best candidate without considering its frequency of occurrence in the corpus.

2. PHONOLOGICAL MAPPING

This paper proposes a hybrid approach that attempts to address both concerns: first, a triphone has to be adequately represented in the corpus in order for a model of it to be well-trained; and second, some triphones have similar phonetic contexts and should be merged together.

2.1. UNDERREPRESENTED AND UNDERTRAINED TRIPHONES

A set of underrepresented triphones is determined from frequency of occurrence of triphones in the corpus. Table 1 shows the number of triphones with the fewest occurrences in the ARPA Resource Management (RM) corpus, including the training set of 3,990 sentences and the test set of 600 sentences (February and October 89).

If one assumes that one occurrence of a triphone is not adequate for training a model of that particular triphone, how many occurrences of a triphone are needed before it is considered to be adequately represented for the purpose of training? The usual strategy is to answer this question by setting a count threshold T empirically and discarding triphones with number of occurrences less than T [2]. In other words, T is the required number of occurrences of a triphone. The solution here is to map these infrequent triphones into phonetically similar triphones rather than sim-

No. of occurrences	No. of triphones
1	368
2	1,322
3	290
4	628
5	356
6	415

Table 1. Triphone frequency in RM.

	Labial	Dental	Alveolar
Nasal	M		N
Fricative	F, V	TH, DH	S, Z
Stop	P, B		T, D

Table 2. Examples of place and manner features.

ply to replace them with their context-independent counterparts.

2.2. DETERMINING PHONETIC SIMILARITY

Once the set of underrepresented triphones and the set of adequately represented triphones are determined, the former is mapped into the latter. In this map, each triphone in the domain can potentially map into one of a number of different triphones in the range. A series of steps are taken to determine the best candidate for the mapping.

First, all phones are categorized according to traditional features involving places and manner of articulation. For each underrepresented triphone, the potential candidates are ranked. The best candidate has a left or right context identical to that of the underrepresented triphone, and the other context belonging to the same phonetic class as the corresponding context of this triphone. The second-best candidate has both contexts belonging to the same phonetic class as the corresponding contexts of this triphone. Decisions on what constitutes a phonetic class are based on knowledge of contextual variations due to coarticulation. For instance, the place of articulation of a phone may determine how its adjacent phones vary. Take the contextual variants for the phones /s/ and /z/. When they are adjacent to a labial consonant such as /p/, /f/, or /m/, the lower cutoff in frication energy falls in frequency. Since this frequency-lowering effect occurs regardless of whether the adjacent consonant is /p/, /f/, or /m/, there is no need to construct three variants of /s/ and three variants of /z/ depending on whether its neighbor is /p/, /f/, or /m/, but instead, the three contexts can be merged into one, since they all share the place feature [+labial]. Conversely, dental fricatives have a frequency-raising effect on /s/ and /z/, and the feature [+dental] is considered a phonetic class for this purpose. It is interesting to note here that even though the phonotactic constraints and phonology of English eliminate many of the possible phone sequences, they are all possible when interword phone sequences are taken into account. Table 2 shows examples of place and manner features. If this ranking of phonetic similarities yields more than one potential candidate, the one that occurs most frequently is chosen.

	IY	AO	P	T	K
Vocalic	2	2	0	0	0
Consonantal	0	0	2	2	2
High	2	0	0	0	2
Low	0	2	0	0	0
Back	-1	2	-1	-1	2
Anterior	-1	-1	2	2	-1
Coronal	-1	-1	-1	2	-1
Round	-1	2	-1	-1	-1
Tense	1	0	0	0	0
Voice	1	1	0	0	0
Continuant	1	1	0	0	0
Nasal	-1	-1	-1	-1	-1
Strident	-1	-1	-1	-1	-1
Labial	-1	-1	2	-1	-1

Table 3. Examples of phonological feature representation.

Second, if a good candidate has not been found in the preceding step, further attempts are made to find the best candidate among the triphones in the range by first representing each context phone as a bundle of phonological features, in the form of a feature vector. Then the problem becomes one of determining which triphone in the range has context feature vectors that are most similar to those of the triphone to be mapped. The features used here are basically Chomsky-Halle features [3], with the addition of [labial]. The values of the features have been changed from binary to values of {-1, 0, 1, 2} in order to assign different weights to different features depending on their different roles in causing contextual variations. In brief, two possible values are available for the presence of each feature, and two for the absence of each feature, depending on how strongly present or how strongly absent (or contradictory) the feature is for a particular phone. Table 3 shows examples of the phonological feature representation.

In the case of diphthongs, feature vectors are assigned to only the left half of the diphthong when it serves as the right context of a triphone, and vice versa. For examples, the dipthong /aw/ has the features of /w/ when it is the left context of a triphone and the features of /a/ when it is the right context of a triphone. Similarly, the affricate /ts/ is analyzed into its component parts: as left context it is the same as /s/, and as right context it is the same as /t/.

To determine how similar a feature vector of a phone in the domain is to a feature vector of a phone in the range, the dot product is used. The assignment of the four values for the features is an attempt to maximize the dot product for phonetically similar triphones and minimize the dot product for triphones that are maximally different from each other. The best candidate for the mapping is the one with the largest sum of dot products for the left and right contexts as well as the highest frequency of occurrence.

The following are three examples of phonological mappings:

```
AE(B,NG) is mapped into AE(P,K)
  Not chosen:
  {AE(M,K), AE(D,G), AE(B,JH), AE(B,D)...}

S(TH,G)e is mapped into S(T,K)b
  Not chosen:
  {S(T,AH)b, S(T,AY)b, S(D,AH)b, S(T,AA)b...}

S(TH,K)e is mapped into S(T,K)b
```

The triphones are notated with the base phone followed by an opening parenthesis, its left and right context phones, and closing parenthesis; "b" or "e" after the closing parenthesis indicates that it is an interword triphone, occurring at the beginning of a word in the former case, and at the end of a word in the latter. The triphones considered in the example mappings but not chosen are shown in the ranked order determined by the mapping algorithm.

The mapping algorithm is summarized as follows:

1. Place all triphones with less than T occurences in the domain, and place all triphones with T occurrences or more in the range.

2. For each triphone in the domain ($triphone_d$), examine in the range each triphone that has the same base phone ($triphone_r$):

2.1. If $triphone_r$ and $triphone_d$ have the same left context, and their right contexts are in the same phonetic class, then include $triphone_r$ in the set of potential candidates for $triphone_d$.

2.2. If $triphone_r$ and $triphone_d$ have the same right context, and their left contexts are in the same phonetic class, then include $triphone_r$ in the set of potential candidates for $triphone_d$.

2.3. If the set of potential candidates for $triphone_d$ is not empty, then go to step 3. Otherwise:

2.3.1. If the left contexts of $triphone_r$ and $triphone_d$ are in the same phonetic class, and if the right contexts of $triphone_r$ and $triphone_d$ are in the same phonetic class, then include $triphone_r$ in the set of potential candidates for $triphone_d$.

2.4. If the set of potential candidates for $triphone_d$ is not empty, then go to step 3. Otherwise:

2.4.1. For each $triphone_r$, compute the dot product of the left context feature vectors of $triphone_d$ and $triphone_r$ (dp_l), the dot product of the right context feature vectors of $triphone_d$ and $triphone_r$ (dp_r), and the sum of dot products (dp_{sum}). If $triphone_r$ has the maximum dp_{sum}, include it in the set of potential candidates for $triphone_d$.

3. From the set of potential candidates for $triphone_d$, choose the one with the largest number of occurrences.

3. EVALUATION

To determine the effects of phonological mapping, a series of tests has been conducted with a version of the SPHINX-II system, which makes use of shared-distribution modeling [4, 5]. For the RM task, there are about 7,700 triphones (including interword triphones), which are clustered into 4,500 shared distributions using one-codebook discrete HMMs. With this shared distribution map, four-codebook semicontinuous HMMs are trained by the forward-backward algorithm. The proposed mapping algorithm is evaluated with different values for the count threshold T, resulting in different sets of phonological mappings. For each phonological map, retraining is carried out with a reduced set of triphones clustered into different shared distributions. The recognition tests are done with a word-pair grammar, and the results are shown in Table 4. The baseline case, which uses all possible triphones (i.e., T = 1) and no phonological mappings, has a word error rate of 3.5%. When phonological mapping is used, the largest word error rate reduction is achieved by setting T to four—i.e., all triphones with number of occurrences less than four are mapped to other triphones that are determined to be phonetically similar. When the domain of mapping is determined only by the value of T being equal to four, the word error rate is reduced by 13.1%. When the domain also includes the eight unseen triphones (i.e., triphones in the test set but not in the training set), the word error rate is reduced by 17.5%.

T	Error rate	Error reduction	Mappings	Undertrained
3	3.1%	10.9% (-1.2%)	1,692	35
4	3.0%	13.1% (1.2%)	1,982	21
4*	2.9%	17.5% (6.2%)	1,990	15
5	3.1%	10.9% (-1.2%)	2,610	14

Table 4. Effects of phonological mapping on error rates, after retraining. (*Including 8 unseen triphones.)

Table 4 also shows the number of mappings, or the size of the mapping domain, for each value of T, as well as the number of undertrained triphones that still remain. Here a triphone model is considered undertrained if one or more of its HMM states fail to be adequately estimated due to insufficient data. For most of the triphones in question, all five of the states are problematic. For the baseline case with T = 1 and no mapping, there are 339 undertrained triphones. The results show that using phonological mapping may reduce the undertrained triphones by up to 96%.

Table 5 shows recognition results using the same phonological maps but without retraining, which means essentially that the underrepresented triphones are simply discarded from the training data. Compared to the baseline case employing all possible triphones, there is a small but still significant reduction in error rate. Since varying T does not make any significant difference, this approach appears to have the most impact when triphones with similar contexts are merged before the clustering of HMM states. Apparently this merger of triphones helps to ensure adequate estimation of the large number of parameters required.

Another baseline for comparison is the best result achieved with no phonological mappings: a word error rate of 3.1% when T is set to three (i.e., all triphones with number of occurrences less than three are replaced by their context-independent phones). In Table 4, the second percentage in parentheses under "Error reduction" is the percentage of error reduction if the results are compared to this higher baseline. In this case, the largest reduction in word error rate is 6.2%. This shows that it is better to replace an undertrained or unseen triphone with a phonetically similar triphone than with its context-independent counterpart.

T	Error rate	Error reduction
3	3.2%	8.2%
4	3.2%	7.7%
4*	3.2%	8.2%
5	3.2%	7.1%

Table 5. Effects of phonological mapping on error rates, without retraining.

4. CONCLUSION

This paper has proposed a general method for reducing the number of triphones in any large vocabulary task, and in doing so has made an attempt to confront the issue of trainability versus specificity in the modeling of subword units. The mapping algorithm takes triphones that are underrepresented in the corpus (and are therefore likely to be undertrained) as the domain and maps each of them into the range consisting of the set of triphones well-represented in the corpus (and are therefore likely to be well-trained). This merger of triphones that have similar phonetic contexts improves the subsequent clustering of shared distributions and the parameter estimation by the forward-backward algorithm, as shown by the test results. It is a completely general method, as it can map unseen triphones in exactly the same manner and in fact can potentially map any set of triphones to a smaller subset. As our understanding of contextual phonetic variations improves, this proposal can also serve as a flexible framework that can easily incorporate such knowledge and thereby improve the modeling of subword units.

REFERENCES

[1] Hwang, M., Huang, X. and Alleva, F. (1993). Predicting unseen triphones with senones. *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, pp. II-311-314.

[2] Rabiner, L. and Juang, B. (1993) *Fundamentals of Speech Recognition.* Englewood Cliffs: Prentice Hall.

[3] Chomsky, N. and Halle, M. (1968) *The Sound Pattern of English.* New York: Harper and Row.

[4] Hwang, M. and Huang, X. (1992). Subphonetic modeling with Markov states models. *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, pp. 33-36.

[5] Huang, X., Alleva, F., Hon, H., Hwang, M., Lee, K. and Rosenfeld, R. (1993). The SPHINX-II speech recognition system: an overview. *Computer Speech and Language* 7(2), pp. 137-148.