A Study on the Triphone Replacement in a Speech Recognition System with DMS Phoneme Models

*Gang-Seong Lee

※ The present research has been conducted by the Research Grant of Kwangwoon University in 1998

ABSTRACT

This paper proposes methods that replace a missing triphone with a new one selected or created by existing triphones, and compares the results. The recognition system uses DMS (Dynamic Multisection) model for acoustic modeling. DMS is one of the statistical recognition techniques proper to a small - or mid - size vocabulary system, while HMM (Hidden Markov Model) is a probabilistic technique suitable for a middle or large system. Accordingly, it is reasonable to use an effective algorithm that is proper to DMS, rather than using a complicated method like a polyphone clustering technique employed in HMM-based systems.

In this paper, four methods of filling missing triphones are presented. The result shows that a proposed replacing algorithm works almost as well as if all the necessary triphones existed. The experiments are performed on the 500+ word DMS speech recognizer.

I. Introduction

The phone is the acoustic model unit generally being used in the speech recognizer. This small unit requires just a little memory, compared with the other larger units such as syllable or word, but it varies its acoustic feature considerably depending on the environment with which the phone is surrounded. Even though the number of context-independent phones (monophones) is not large, as we consider a phone surrounded by different phones, the number of context-dependent phones becomes big. This kind of context-dependent phone is also called 'polyphone'. Many recognizers use polyphones, and especially triphones, which are basically phones observed in the context of given preceding and succeeding phones (± 1), and have been the most common for many years[1]. The polyphone is still one single phone which is modeled depending on its context. There are wider context-dependent acoustic models than the triphone. If two phones are considered on each side of a center phone (± 2), it becomes quintphone.

Finke and Rogina[2] in the WSJ(Wall Street Journal) data base show how many different models are obtained when using different context sizes. The number of models of context-width 1 (triphone) was approximately 35,000, context-width 2 (quintphone) was 130,000 and context-width 3 was 160,000. These figures are not always the same as when the another text corpus is used, but it gives us an estimation of how many different models can exist.

Having a complete set of polyphones is almost impossible. When a new word is registered, acoustic models for the phones in the word should be found in a polyphone database. If there is no exact model, then the closest model is substituted for the missing polyphone. One possible way to solve this problem in HMM (Hidden Markov Model) is to use a decision tree[3-4]. All the polyphones sharing a common center phone are clustered into a tree and each leaf has its own weight vector corresponding to the shared code book. This is an effective way to reduce the number of polyphones and find a good replacement.

Another technique is required for different acoustic modeling. The system described in this paper is based on DMS (Dynamic Multisection) model [5], which is different from HMM, and context clustering tree technique is not applicable directly to DMS. The system is designed for the isolated-word recognition of 500+ words using triphones[6]. Because of the native features of DMS, each triphone has its own code book; there are no weight vectors as in HMM, and, further, DMS has a length information in each section, which makes triphones hard to share their codebooks. If a new word is registered in a dictionary and a new triphone is required that is not registered in acoustic models, the system has to prepare the triphone by either replacing it with an existing one or creating a

* Computer Engineering Dept.Kwangwoon University
Manuscript Received : April 23, 1999.

new one. In this paper, four different methods of preparing new triphone are presented and experimental results are shown. The recognition system is described briefly in section 2, replacement methods in section 3 and the experimental results are shown in section 4.

II. Dms-Based Recognition System

The system is designed originally for real time, isolated/-continuous word speech recognition using DMS triphone acoustic models. In this paper, however, only vocabulary-independent isolated speech recognition part is described. The overall system block diagram is shown in Fig 1.

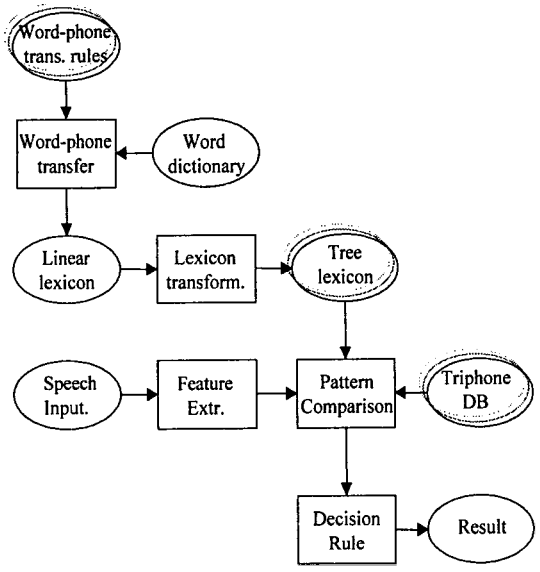


Figure 1. Block diagram of recognition system.

In the recognition phase, the recognizer uses two major data base sources - triphone DB and tree-structured lexicon. The triphone DB is a collection of triphones that are placed linearly in a file. In the tree-structured lexicon, two or more words which contain the same initial phones, share a single sequence of phone models. The existing tree search algorithm to find a word in a tree-structured lexicon is quite efficient to speed up the recognition process, and it is used on a widespread basis in the main decoding process[7].

The flat-structured lexicon, which is a basic form of the tree-structured lexicon, is produced by 'word-to-phone transformer' taking an input of a ASCII format file, 'word-dictionary' together with the file, 'word-phone transform rules'. The only thing a user has to do to initialize the recognizer is to create a word-dictionary.

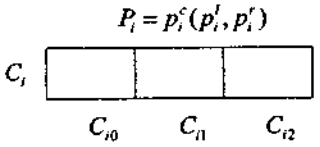
When an input pattern is compared with words in the lexicon, the tree-structured lexicon requires specific type of triphones, which should be found in triphone DB. All the possible triphones, however, are not registered in the triphone database; when a triphone is required, a replacement technique is applied here and it produces a replaceable triphone that is the closest to the required one. A detailed description of these replacement techniques is given in the next section.

III. Replacement Methods

One triphone model consists of three different sections (or states) in this system, which possibly represent left, center and right parts of a phone depending on the context. These triphones are stored in triphone data base (Fig 2). Suppose B is a complete set of phones, $p_i^c, p_i^l, p_i^r \in B$. Triphone $P_i = p_i^c(p_i^l, p_i^r)$ represents a center phone p_i^c, which is surrounded by a left phone p_i^l and a right phone p_i^r. For instance, a phone sequence of /# a b c d #/ can be represented by a triphone sequence of /a(#,b) b(a,c) c(b,d) d(c,#)/, where # is a silence.

Triphone name	Model
$P_1 = p_1^c(p_1^l, p_1^r)$	acoustic model of $p_1^c(p_1^l, p_1^r)$
$P_2 = p_2^c(p_2^l, p_2^r)$	acoustic model of $p_2^c(p_2^l, p_2^r)$
...	...
$P_i = p_i^c(p_i^l, p_i^r)$	acoustic model of $p_i^c(p_i^l, p_i^r)$
...	...
$P_M = p_M^c(p_M^l, p_M^r)$	acoustic model of $p_M^c(p_M^l, p_M^r)$

(a)



(b)

Figure 2. Triphone data base structure (a) and structure of each triphone (b) Each triphone has three sections and each section has one vector.

Method 1 - Using PDT

This is a simple algorithm. A required triphone is taken from one of the existing triphones. The algorithm selects one triphone that is the closest to the required one, and replaces it. To do this, we have to know the distance between two triphones. However, it is not

possible to get the acoustic distance between required triphone and existing one, because the required one is not registered. We only can get the distance of phonemic symbols. To do this, a matrix called *PDT*(Phone Distance Table) is used. This table contains pseudo-acoustic-distances between phones (Fig 3). The distance is determined logically under the consideration of phonetics. For instance, the distance between completely different groups of phonemes - like plosives and vowels - is set to 20 and between phonemes within same group - like weak explosives(P, T) - to 1. This might not be the best way to determine the distances, but it is one of several possible ways.

The distance $d_P(P_i, X)$ between two triphones (P_i, X) is obtained by simply looking up the table *PDT*.

$$d_P(P_i, X) = d_T(p_i^c, x^c) + d_T(p_i^l, x^l) + d_T(p_i^r, x^r)$$

where, $d_T(a, b)$ is the value taken from the table *PDT* - row a and column b.

The closest triphone Y is obtained using the following equation.

$$Y = \underset{P_i}{minarg}\ d_P(P_i, X)$$

	K	N	T	···	a	e	i	···
K	0							
N	20	0						
T	1	20	0					
···				···				
a	20	20	20	···	0			
e	20	20	20	···	2	0		
i	20	20	20	···	5	5	0	
···								···

Figure 3. PDT (Phone Distance Table).

Method 2 - Replacing each section with a center section of other triphone

Lets suppose that, in the first method, the required triphone $X = a(b, c)$ is replaced by $X' = a(b, d)$, where a, b, c, d $\in B$. In this case, it is logical to suppose that right part (third section) of triphone X' is not proper to the triphone X. By replacing the right part (third section) of the acoustic model X' with a value closer to phone c, we will be able to get a better acoustic model of X'. Likewise, the same logic is applied to center and left phones. Here is the algorithm.

Required triphone : $X = x^c(x^l, x^r)$

Triphone replaced by method 1 : $Y = y^c(y^l, y^r)$

if $x^l \neq y^l$:

Get $S = x^l(\#, \#)$ or get the closest triphone S from method 1

replace first section of acoustic model Y with center section of acoustic model S

if $x^c \neq y^c$:

Get $S = x^c(\#, \#)$ or get the closest triphone S from method 1

replace center section of acoustic model Y with center section of acoustic model S

if $x^r \neq y^r$:

Get $S = x^r(\#, \#)$ or get the closest triphone S from method 1

replace last section of acoustic model Y with center section of acoustic model S

Method 3 - context-independent phone

This is an averaging technique. It takes all the triphones which have the same center phone as the required one(x^c), to make a normalized model - this is quite the same as the context independent model.

$$Y = \frac{1}{N} \sum_{p^c = x^c} P$$

where, N is the number of P which is $p^c = x^c$.

Method 4 - separate calculation

Each section of the acoustic model is calculated separately and collected together. For left section, all the triphones which have the same context as $x_l x_c$ on the left are accumulated and normalized to get L, and the first section of L is taken for the first section of Y. Likewise, for the center and right section, triphones which have the context of x_c and $x_c x_r$ are averaged to get center and right section of Y. In this way, we get a new triphone acoustic model Y which represents local context well.

$$L = \frac{1}{N_l} \sum_{p_l = x_l,\ p_c = x_c} P$$

$$C = \frac{1}{N_c} \sum_{p_c = x_c} P$$

$$R=\frac{1}{N_Y}\sum_{p_c=x_c,p_r=x_r}P$$

where, N_l is the number of P which is $p_l=x_l$ and $p_c=x_c$, and so on for N_c and N_r. Y takes left section from L, center section from C and right section from R.

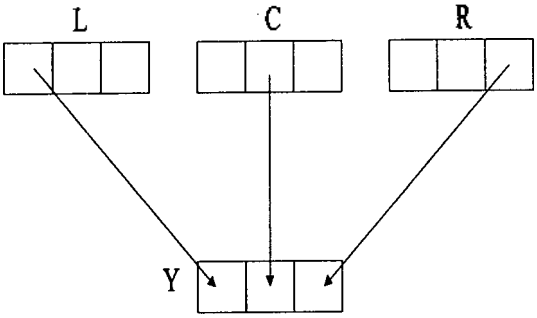


Figure 4. Separate calculation of each section.

IV. Experiments

4.1. Signal Processing

Input speech was sampled at the rate of 10kHz with 16 bits resolution. LPC cepstrum coefficients of order 10 were extracted for a feature parameter. A frame size is 128 samples and adjacent frames are separated by 128 samples.

4.2. Data base

Five hundred sixty-one (561) district names of Seoul and Pusan were recorded three times by two male speakers. First recording was used for training and two repetitions were used for baseline testing which doesn't need any triphone replacement. Another set of 120 human names and department names of a university was recorded twice by two same male speakers. These utterances were used only for testing.

4.3. Training

First repetition of 561 words is used to make a triphone data base. The number of triphones extracted from the set was 1103, which is not a complete set of triphones. Each triphone has three sections and each section has one codeword.

4.4. Testing

First testing (S-120) is to see how good the baseline performance is, when all the necessary triphones are in the triphone data base. Two repetitions of 120 words - a subset of 561 words used for training - of two male speakers are tested and give 97.08% of recognition result. When the vocabulary size is increased to 561 (S-561) - full set, recognition rate is 95.01%.

Second testing is based on the vocabulary of names of people and departments, which is a totally new set of 120 word vocabulary. Among the required 1129 triphones, 290 triphones are not found in the data base (25.7% missing). Table 1 shows the results of each test. M1 means method 1 described in section 3, etc.

Table 4. Recognition Results.

	S-120	S-561	M1	M2	M3	M4
Recognition Rate	97.08%	95.01%	86.67%	91.67%	92.50%	96.67%

V. Discussion

The recognition rate for the vocabulary set of 120 words with a complete triphone set is the best - 97.08%. As the size of vocabulary increases, the performance is expected to become lower. For the test vocabulary set of 561 words, the recognition rate is 95.01%. Statistical methods (M3, M4) were always better than the other simple replacement methods (M1, M2). When contextual information is considered (M2, M4), it gives better results than when it is not considered(M1, M3). Method 4 turned out to be the most efficient way to replace missing DMS triphones among four algorithms, and the performance difference is only 0.41% in this case, that is almost comparable to the result when no triphone is missing.

VI. Conclusion

Four different methods which replace a missing triphone with a new one selected or created by existing triphones are proposed. The best result gave 96.67% of recognition rate when the vocabulary which misses 25.7% of triphone is used. That performance is only 0.41% lower than the vocabulary set which has complete triphone models. This demonstrates that the proposed method M4 is very effective replacement method in DMS recognition system.

References

1. Lee, K. Context-Dependent Phonetic Hidden Markov Models for Continuous Speech Recognition. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Apr. 1990, pp. 599-609.

2. Michael Finke, Ivica Rogina "Wide context acoustic modeling in read vs. spontaneous speech," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997.
3. Hwang, M.Y. "Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition," Ph.D. Thesis, Carnegie Mellon University, 1993.
4. Ivica Rogina, "Automatic Architecture Design by Likelihood-Based Context Clustering With CrossValidation," *Proceedings of Eurospeech-97*, 1997.
5. Gang Seong Lee, "Speaker Independent Isolated-Word Speech Recognition Using DMS," *Journal of the Institute of New Technology in Kwangwoon Univ.*, Vol. 26, 1997.
6. Gang Seong Lee, "500+ Words Isolated-word Speech Recognition System," Acoustic Society Korea Conference on Acoustic, Speech and Signal Processing, pp. 83-86, Vol. 17, No. 1(s), 1998.
7. Mosur K. Ravishankar, "Efficient Algorithms for Speech Recognition," Ph.D. Thesis, Carnegie Mellon University, 1996.

▲Lee Gang Seong

Lee, Gang Seong was born on January 15, 1964 in Seoul, Korea. He received the B. S. degree in computer engineering, and the M.E. and Dr. Eng. degrees in the same field from Kwangwoon University, Seoul, Korea, in 1986, 1988, and 1993, respectively.

In 1991, he joined the faculty of Computer Engineering Dept in Kwangwoon University, and he has been an Associate Professor since 1998. He studied at Carnegie Mellon University since Aug. 1998, for a year, as a visiting research scholar at School of Computer Science. His research interests are speech recognition and language modelling.