音声言語情報処理 35－7
（2001. 2. 2）

MODELING SEGMENTAL CONTEXT EFFECT FOR CHINESE SPEECH RECOGNITION

Jin-Song Zhang, Shu-Wu Zhang, Yoshinori Sagisaka, Satoshi Nakamura
ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-0288 Japan
{jzhang, szhang, sagisaka, nakamura }@slt.atr.oc.jp

ABSTRACT

This paper reports our approach of modeling the segmental coarticulatory effects for Chinese speech recognition. When building triphones based on the Initials and Finals, the number of possible triphone units are very large and easily lead to serious unseen phone problem. We propose to cluster the segmental context effects into a few categories according to articulatory configurations in order to reduce the number of triphone units, and do phonetic-decision-tree (PDT) based state tying at the level of category to improve data-sharing when training HMMs. The proposal is able to reduce an original 111k possible triphones to 4.6k ones and nearly solve the unseen triphone problem. Application to both a task-dependent(TDEP) and a task-independent(TIND) experiments of speaker independent continuous speech recognition showed that the proposal brought by relative 14% error reduction in the TIND task, and had no significant differences in the TDEP task.

Keywords： coarticulation, articulatory configuration, unseen phones, task-dependent speech recognition, task-independent speech recognition, Chinese speech recognition.

中国語音声認識における音素環境のモデリング

張勁松、張樹午、匂坂芳典 、中村哲
ATR 音声言語通信研究所
〒 619-0288 京都府相楽群精華町光台 2-2-2
{jzhang, szhang, sagisaka, nakamura }@slt.atr.oc.jp

あらまし

トライフォンモデルを中国語音声認識に応用する場合、出現するモデルの数が膨大となり、モデルの学習を十分に行うことは難しい。そのため、タスクの変更により音素環境のミスマッチが生じた場合、認識性能は大きく劣化する。そこで我々は音響環境を考慮し、調音特性をもとに音素環境を分類するとともに、この分類に対し適当な調音質問による文脈分類木を作成した。これにより音響モデルをデータ駆動クラスタリングにより状態共有することで、モデル数を大幅に削減することが可能となり、タスクの変更に伴う音素環境のミスマッチに関する問題を解決することができた。タスク依存データ・非依存データを用いた認識実験の結果、タスク非依存データにおいて１４％の誤り改善率を得るとともに、タスク依存データに対して性能の劣化が無いことを確認した。

キーワード： 音素環境、調音特性、中国語音声認識、未学習音響環境、タスク依存認識、タスク非依存認識。

1. INTRODUCTION

As a consequence of the fact that neighboring sounds have the most influences on a sound among a number of sources of variability in the acoustic properties of speech[1], context-dependent (CD) acoustic model plays an important role in the state-of-art automatic speech recognition systems. In this paper, we discuss the problems in the application of CD hidden Markov models (HMMs) to Mandarin Chinese speech recognition, introduce our segmental context clustering method which is based on articulatory configuration, and report speech recognition results of two tasks: a task-dependent(TDEP) testing with simlar phonetic environments to the training data, and a task-independent (TIND) testing with a different phonetic environments from the training data. [2].

Among various kinds of context-dependent strategy, triphone modeling is to develop a phone acoustic model dependent on both the preceding and the succeeding phones, hence clearly model the possible forward and backward coarticulatory effects. It has shown robust performance gains and turns out the standard approach in the literature. However, one well-known accompanying problem of this technique is the sparse data problem during the development of triphone models. Not surprisingly, the same problem also happens when the triphone modeling technique is applied to Mandarin Chinese speech recognition. Furthermore, compared with alphabetic languages, a specific problem may arise for Chinese due to its monosyllabic morpheme characteristic. Coarticulations of intra-syllabic phones are rather different from those of inter-syllabic phones. Simultaneous modeling of the syllabic boundary information results in a more number of tri-phone models.

One efficient solution to the sparse data problem is data-sharing for the phones with similar phonetic context[7]. Phonetic-decision-tree(PDT) [3, 4] based context clustering is the most popular data-driven method for training CD HMMs. It is also regarded as one good method to solve the problem of unseen CD triphones. However, one easily neglected problem is that the data-driven nature of the PDT-based unseen phone synthesis is less likely to generate robust models for unseen CD phones as there is no data available. Although this problem does not affect the performance of a TDEP recognition, it probably take effects in a TIND recognition where some necessary phones are less robustly synthesized.

In order to deal with the above-mentioned problems for Chinese speech recognition, we proposed to cluster coarticulatory effects into a few categories according to articulatory configurations of speech segments. When we use Initials and Finals as the basic units, the proposal is able to reduce the originally possible 111k kinds of triphones to 4.6k ones (reduced by 24 times). The clustering of context effects are assumed to do least harm to model accuracy. Preliminary experiments on a TDEP and a TIND testings showed the effectiveness of the proposal.

2. CD IF MODELS

A Chinese word is composed of one to several characters, and each character is pronounced as a monosyllable with a pitch tone. The totally phonetically differentiable tonal syllables are about 1,300, and the number of base syllables is about 410 when pitch tones are discarded. Traditional Chinese phonology [5] divides the syllable into demi-syllabic units: an *Initial*(Table1) and a *Final* (Table 2).

We use a basic acoustic unit set consisting of 21 Initials and 37 Finals (IFs) in our speech recogntion system. The 21 Initials are those in Table 1, and the 37 Finals came from the 39 Finals in Table 2 with the /e1,e2/ pair and /ong, ueng/ pair merged. Since /e1/ and /e2/ are dependent on the preceding Initials, and can be discriminated by the CD IFs models after the symbol sharing. /ong/ and /ueng/ are regarded acoustically similar by the authors. The number of base syllable is 412.

2.1. Too large a number of CD IFs models

With the set of 21 Initials and 37 Finals and a /sil/ for silence, we found that the possible number of triphone models are as large as 111,625, including triphone Initials and Finals. The possible number of triphone Initials is calculated as:

$$(37 \text{ Final} + 1 \text{ sil}) \times (412 - 35 \text{ null-Initial}) = 14,326$$

The count for the triphone Finals is a little complicated as it needs to take consideration of tri-syllabic combinations in the case of null-initial syllables.

$$(412 - 35) \times (21 \text{ Initial} + 1 \text{ sil} + 35 \text{ null-Initial}) + (37 \text{ Final} + 1 \text{ sil}) \times 35 \times 57 = 97,299$$

It is impossible to collect enough data for robust estimation for all these models separatively.

2.2. Potential problem of PDT training

Phonetic-decision (PDT) based data-driven model clustering is widely used to solve the sparse data problem in model training[7, 3, 4]. It is regarded as one good method to solve the problem of unseen CD models by synthesizing the unseen CD phones through traversing a binary tree according to the phonetic question in each node. However, the robustness of the synthesis is not guaranteed. As illustrated in Figure 1, the centre state of the unseen CD Initial /a-s+u/ is synthesized using the parameters from /a-s+i/. This result is unwanted in that the /s/ in /a-s+u/ is probably produced with lip-rounding due to the articulatory anticipation for the succeeding /u/, whereas the /s/ in /a-s+i/ is not associated with the configuration. It may be

Manner / Place	Stops		Affricative		Fricative		Nasal	Liquid
	Unaspirated	Aspirated	Unaspirated	Aspirated	Unvoiced	Voiced	Voiced	Voiced
Bilabial	b	p					m	
Labio-dental					f			
Dental			z	c	s			
Alveolar	d	t					n	l
Retroflex			zh	ch	sh	r		
Palatal			j	q	x			
Velar	g	k			h		ng	

Table 1: Pinyin symbols for consonants in Mandarin Chinese, by traditional articulatory categories. All consonants except for /ng/ can serve as Initials.

Final-onset based Classification	Final-offset based Classification								
	a	o	e	i	u	v	n	ng	retro
a	a	ao		ai			an	ang	
o		o			ou			ong	
e			e (e1, e2)	ei			en	eng	er
i	ia	iao	ie	i (i1, i2,i3)	iu		in, ian	ing, iang, iong	
u	ua	uo		ui, uai	u		un, uan	uang, ueng	
v			ve			v	vn, van		

Table 2: Pinyin symbols for Finals of Mandarin Chinese syllables. The table is organized with rows for Final-onset Pinyin symbols and columns for Final-offset ones.

more preferable to synthesize the state for /a-s+u/ using that from /u-s+u/, in which the /s/ similarly has lip-rounding. Since the tree is developed based on the data of seen CD phones and a number of phonetic questions, the variability in the speech data and multiple choices of phonetic questions easily lead to different tree structures in different situations such as different training data or different phonetic questions used. As a result, consistent and robust syntheses of unseen CD phones cannot be guaranteed in all cases.

Besides, [2] proposed an efficient unseen triphone synthesis method by tying the first and the last state of the desired triphone with the first state of the corresponding left bigram CD model and the last state of the corresponding right CD model respectively. But if we pay a glance at the PDT tying results of CD triphones, we may see frequent references to the right context for tying the first state, and frequent references to the left context for the third state. It indicates that training triphones from data is prefered as they generally better model the coarticulation than those synthesized based on biphones, if the unseen phone problem can be solved by other ways.

3. CLUSTERING SEGMENTAL COARTICULATORY EFFECTS

Another solution to the sparse data problem is to cluster the coarticulatory effects into a few categories according to human knowledge when developing CD models. For example, two stop sounds /b/ and /p/ are known to arouse mostly same formant transitions to the preceding and succeeding vowels [1], combining their coarticulatory effects into one kind can nearly improve the data-sharing by a double times without decreasing the modeling accuracy. There are some previous studies based on this idea for Chinese speech recognition. For example, [8] suggested to categorize 38 Finals into 12 groups based on their ending phoneme and categorize the 22 Initials into 7 to 11 groups based on articulation phenomena when developing inter-syllabic CD models.

One defect of these preliminary studies was that the approaches were separated from data driven method to do model clustering. Moreover, the testing was TDEP task. Since there were no robust gains against data-driven methods, the method was gradually abandoned.

However, the method offers a potential high-degree data-sharing which is a desiring answer to reduce the over than 111k triphones. Also, the high-degree data-sharing should probably lessen much the problem of unseen tri-phone IFs in TIND task. Therefore we revive the method and further adopt the phonetic decision tree model to do data-driven context-effect clustering on the level of category.

What is important now is how to categorize the Initials and Finals, and how to design appropriate phonetic questions for the context categories. We did the categorization according to the articulatory configurations of each Initial and Final. Acoustic theory revealed that the formant transitions are a

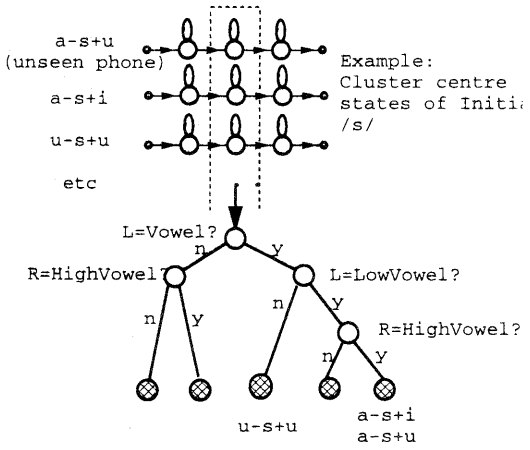

Figure 1: Illustration of decision tree-based synthesis of unseen phone state.

cue to place of articulation and are not restricted to any given manner of production[1]. For example, the nasal consonant /m/ has the similar formant transitions to the stop consonants /b/ and /p/ as it also starts from a bilabial stop occlusion. Therefore the coarticulatory effects of the 21 Initials are classified into 7 groups: labial, labial-dental, dental, alveolar, palatal, retroflex and velar, just like the rows in Table 1.

A	a, ia, ua
E	e, ie, ve
I1	ai, ei, uai, ui
I2	i1, i2, i3
V	v
U	u, ou, iu,ao,iao
O	o, uo
R	er
G	iang, ing, uang, ong, eng, ang, iong
N	ian, in, uan, un, en, an, van, vn

Table 3: Final categories for carryover coartilatory effects classification.

The compound Finals have different onset and offset configurations, thus there should be different categorizations of a Final with respect to the onset and offset. Table 3 gives our categorization of carryover effect based on the Final offsets, and Table 4 for anticipatory effect based on the Final onset. A special consideration of glottal stop is paid to the null-Initial syllables. Speech synthesis study has found that negligence of the glottal stops let the synthetic output for sequences such as 一五一十 (yiwuyishi) difficult to be understood[6]. Therefore, we adopt 9 symbols to represent the glottal stops which serve as the anticipatory effects from null-Initial syllables.

I	hI	i1, i2, i3, ia, ian, iang, iao, ie, in, ing, iu, iong
U	hU	u, ua, uai, uan, uang, ui, un, uo, ong
V	hV	v, van, ve, vn
E1	hE1	e, er
E2	hE2	ei,
E3	hE2	en
E4	hE2	eng
O1	hO1	o
O2	hO2	ou
A1	hA1	a
A2	hA2	ai, an
A3	hA2	ao, ang

Table 4: Final categories for anticipatory coartilatory effects classification. The second column indicates the symbols for glottal stop.

After we decompose each syllable into a right CD Initial and a left CD Final, we got 187 right CD Initials and 152 left CD Finals from the 412 base syllable list. From them, we can estimate the possible category-based (CB) triphone Initials:

$$(10 \text{ Final carryover effects} + 1 \text{ sil}) \times (187 \text{ right CD Initial}) = 2,057$$

and the count for the number of CB triphone Finals

$$152 \text{ left CD Finals} \times (7 \text{ Initial classes} + 9 \text{ Final glottal stops} + 1 \text{ sil}) = 2,584$$

The total number is 4,641 , which is only 4.1% of the original 111,625 kinds.

Phonetic decision tree is used to do data-drive HMM state tying on the level of categories. The questions asked for the Initial categories are designed according to the constriction place in the midsagittal plane and tongue movements. For example, labial-dental includes the two categories labial and dental, alveolar-palatal for both alveolar and palatal ones, tongue-blade for palatal and retroflex ones and etc. The questions asked for the Final categories are much like the normal ones for the phone targets, such as high-vowel, low-vowel, lip-rounding, and etc.

4. EXPERIMENTS AND RESULTS

4.1. Experimental setup

Our baseline system is a speaker independent, intra-word triphone, decision-tree based tied-state system. Each triphone IF is modeled by a left-to-right, three-emitting-state continuous density HMM without state skipping. Silence is modeled by two models: a three-emitting-state HMM and a one-shared-state HMM allowing state skipping.

Input speech, sampled at 16kHz, was initially pre-emphasized $1-0.97z^{-1}$ and grouped into frames of 25ms with a frame shift of 10ms. For each frame, a Hamming window was applied followed by the computation of 12 MFCCs and normalized log-scaled energy. The first order time derivatives were computed and added to the feature. Thus each speech frame was represented by a vector of 26 features.

The vocabulary consists of 9,754 words and the anguage model is a bi-gram word model trained from the mixed data of newspaper transcripts and Hotel-Reservation-task dialogue sentences, totally about 300k words. The recognition performance is evaluated according to the accurate rate of Chinese character (CCR). All the recognition experiments are performed using the search engine of the HTKv2.1 toolkit.

4.2. Training and Testing Data

Training data: The training database is the 863 Putonghua corpus, acquired from mainland China. The corpus consists of a total 166 native speakers, 83 males and 83 females, each reading about 580 sentences with rich phonetic context information. The base syllable appearing at least once is of 401 kinds. The totally uttered syllables amounts to 1,176,148, and the data occupies about 14 Gigabytes disk space.

Task-dependent testing database: Two of the authors, both male native Chinese speakers, each read a different set of 650 sentences from the 863 corpus. The speech is recorded by portable DAT at a sampling rate of 44.8kHz, then downsampled to 16kHz by DATLink. This data is used for the TDEP testing.

Task-independent testing database: The data used for task-independent testing was also collected by the two same speakers. The text set consists of 237 sentences of five Hotel Reservation client-customer dialogues (HRD). Each speaker uttered all the 237 sentences, and the data were recorded at the same condition of collecting TDEP testing data. Since 863 sentences were mainly selected from newspapers, it could be imagined that there exist considerable mismatch of phonetic context in the two kinds of database.

	Least number of samples per unit				
	1	10	20	30	40
863	6417	6417	5611	5432	5352
Percentage 1	5.7%	5.7%	5.0%	4.9%	4.8%
HRD	837	837	799	792	785
Percentage 2	74.5%	75.4%	72.0%	71.4%	70.7%
Percentage 3	84.0%	84.0%	82.4%	81.4%	81.3%

Table 5: Distribution information for triphone IFs in the 863 training data and HRD task.

4.3. Triphone IF coverage

Based on the lexicon, we generated triphone IFs from the 863 transcript and the HRD transcript separately. As a result, there are 6,417 kinds of triphone IFs appearing at least once in the training data, and 1,110 triphone IFs appearing at least once in the HRD transcript. The phonetic context mismatch of the two tasks can be examined by a look at the coverage of HRD triphones in the 863's as a function of the least number of samples per unit in the training data, as shown in Table 5. Table 5 also gives the following information in percentages

- Percentage 1: percentage of the categories of appearing triphones in the training data to the 111,625 possible triphone IFs.
- Percentage 2: percentage of the categoriess of appearing triphones to that of all appeared 1,110 triphones in the HRD transcript.
- Percentage 3: percentage of accumulated phone occurances to the total one in HRD transcript. For example, although only 74.5% triphone categories have at least one sample in the training data, it amounts to 84.0% of all phone occurances in the HRD transcript.

From Table 5, two things are clear: One is that there are 273 (amount to 25.5% of triphone categoriess in the testing set) necessary triphone IFs for HRD task have no one sample at all in the training data. Although phonetic decision-tree based method will synthesize all the necessary unseen phones for recognition, the un-guaranteed robustness of these synthesized models potentially influence the recognition performance. The other one is that totally appearing triphone IFs in the training data is only 5.7% of all the possible triphone IFs, indicating the unseen phone problem potentially severe.

	Least number of samples per unit				
	1	10	20	30	40
863	2427	2427	2270	2223	2196
Percentage 1	52.3%	52.3%	48.9%	47.9%	47.2%
HRD	728	728	717	716	714
Percentage 2	92.3%	92.3%	90.8%	90.7%	90.4%
Percentage 3	96.3%	96.3%	95.7%	95.6%	95.5%

Table 6: Distribution information for CB triphone IFs in the 863 training data and HRD task.

4.4. CB triphone IFs coverage

After we generated within-word category-based (CB) triphone IFs, we got 2427 kinds of triphones appearing at least once in the training data, and 790 ones in the HRD data. The phone coverage is given in Table 6. By comparing this table with Table 5, the effects of the knowledge-based coarticulatory effect clustering are obvious:

- Only 52 CB triphone IFs desired by the testing transcript have no samples in the training data, whereas over than 92% of desired models are seen in the training data. Compared with the original IFs approach, the unseen triphone problem is much lessened.

- The percentages of accumulated phone occurances are much better than those in Table5.
- Over than 90% of desired models have more than 30 samples per unit in the training data, indicating the sparse data training problem is much lessened.
- The totally appeared CB triphone IFs in the training data accounts for 54% of the all the 4567 possible ones, much better than the original situation.

method	TDEP			TIND		
	Spk1	Spk2	Ave.	Spk1	Spk2	Ave.
triphone IF	91.9	90.8	91.3	86.2	87.4	86.8
triphone-class	92.7	90.2	91.4	88.1	87.3	87.2
+PDT	92.8	90.1	91.5	89.4	87.9	88.7

Table 7: Chinese character rates(CER) using PDT-based state-tied triphone IFs, triphone-class IFs, and PDT based state-tied triphone-class IFs.

4.5. Speech recognition results

We developed triphone models by three methods:

- Baseline triphone IF: 9,664 within-word triphones from lexicon, with 2,757 PDT-based tied states and a total of 8,269 Gaussians. Unseen triphones are synthesized by PDT.
- CB Triphone IF: 2,870 within-word CB triphones, with 8,608 untied states and 8,606 Gaussians. Unseen triphones are interpolated from monophone HMMs.
- CB Triphone IF + PDT: with 2,795 PDT-based tied states and 5,588 Gaussians, Unseen triphone-classes are synthesized by PDT.

From the results in Table 7, one can see that,

- The baseline method degraded by 3.5% CCR (or 40% error increase) when applied to TIND task compared with TDEP, indicating that it possibly resulted from the phonetic context mismatch between TIND and training data.
- Both the two CB triphone IF methods improved CCR performances in the TIND task. Since all the three methods use the same language model and lexicon, the improvements should be ascribed to the acoustic models. This indicates that the proposal of coarticulatory effects clustering did lead to more robust triphone HMMs, and was an effective answer to the unseen triphone IF problem.
- Phonetic decision tree based state-tying achieved the best performance with a total 14% error reduction compared to the baseline in the TIND task, indicating further data-driven clustering efficient on the level of categories.
- There are no significant differences for the three methods about the average performances in the TDEP task. And the proposals do not show consistent performances for the two speakers: one improved and the other one slightly degraded. More speakers data is needed to draw a more reliable conclusion.

5. CONCLUSION

Based on coarticulatory configuration analyses, we grouped the 21 Initials into 7 categories, 22 Finals into 10 carryover categories and 12 anticipatory categories. This proposal reduces an original 111k possible triphones to 4.6k ones and serves as an effective answer to the serious unseen phone problem associated with triphone IFs. Although we probably need to test on more speaker's data, the preliminary experiments on two male's data showed that the proposal brought good error reductions in the TIND tasks, and had no big differences in the TDEP task compared to the baseline system. The results also showed that we had successively applied the phonetic-decision-tree based clustering technique to further improve model robustness at the level of phonetic category.

6. REFERENCES

[1] R. D. Kent and Ch. Read, "The acoustic analysis of speech", Singular publishing Group, Inc., 1992.

[2] Ch.-H. Lee and et al, "A study on task-independent subword selection and modeling for speech recognition", ICSLP96, pp.1820-1323.

[3] Q. Huo and B. Ma, "Irrelevant variability normalization in learning HMM state tying from data based on phonetic decision-tree", ICASSP99, pp.577-560.

[4] Sh. Gao, B. Xu and T.-Y. Huang, "Class-triphone acoustic modeling based on decision tree for Mandarin continuous speech recognition", Proc. From ISCSLP 98.

[5] Z.-J. Wu et al, "Xian Dai Han Yu Yu Yin Gai Yao", Sinolingua Press, Beijing 1992.

[6] Ch.-L. Shih and R. Sproat, "Issues in Text-to-Speech Conversion for Mandarin", Language and Speech 97.

[7] Sh. Sagayama, "Phoneme environment clustering for speech recogntion", ICASSP89, pp.397-400.

[8] L.-Sh. Lee, "Voice dictation of Mandarin Chinese", IEEE Signal Processing Magazine, July, 1997, pp.63-101.