Automatic Question Generation for HMM State Tying using a Feature Table

Supphanat Kanokphara and Julie Carson-Berndsen
Department of Computer Science
University College Dublin,
Ireland
{supphanat.kanokphara, julie.berndsen}@ucd.ie

Abstract

Generally, speech recognition systems use context-dependent subword units for recognition. To utilize these units, tree-based state tying is an unavoidable technique. This tying applies linguistic questions to group these units and ties them accordingly. These questions usually require an expensive resource-consumption manual procedure. Even though this problem has been addressed by data-driven automatic question generation systems, the quality of the systems greatly depends on the quality of the corpus. This paper presents a new approach to automatically assign questions using only a simple feature table which has three advantages. First, the requirement for expert knowledge is reduced to only a simple feature table instead of complex questions. Consequently, question extension or modification is effortless. Second, feature tables still preserve a benefit of tree-based state tying, i.e. incorporating linguistic knowledge into questions. Third, feature tables are commonly employed in phonetic studies for all languages. It is therefore a resource which is easy to access.

1. Introduction

Since phone identities are heavily influenced by adjacent phones, context-dependent subword units are widely exploited as basic units for recognition. Unfortunately, a larger subword inventory often results in insufficient speech data for training. One way to solve this problem is to prune out model states depending on data availability. This is where tree-based clustering is frequently suggested (Odell, 1995). The philosophy behind this technique is that model states having similar acoustic properties can be shared and trained from the same data where the acoustic similarities are determined from linguistic questions which have to be carefully defined.

Manual procedures for question generation regularly suffer from human errors and are time-consuming. The automatic question generation using acoustic data (Singh, Raj and Stern, 1999) has been developed in order to support these problems. The system has three advantages over manual procedures. First, speech data are already available in a corpus, while experienced phoneticians are expensive and hard to find for some languages. Second, a computer works faster than a human. Third, an algorithm judgment is consistent while human judgment is not. Nevertheless, manual procedures are still better than the data-driven system because it leads to other problems. First, the system cannot completely estimate unseen context-dependent subword units. Second, incorrect questions may be produced if speech quality is poor or the database is small.

Both manual procedures and data-driven systems have one thing in common, i.e., they require input to generate questions. Input to a data-driven system is acoustic data while input to a manual procedure is expert knowledge. Since manual procedures and automatic systems both have advantages, now the most obvious question is "Is it possible to design an automatic system requiring existent or almost ready input which covers unseen units and is robust to noise?"

A feature-table-based system is one possible answer to this question. Feature tables are commonly used in phonological descriptions of most languages and are thus readily available. Even though there is no feature table existing for some languages, the system limits human intervention to only feature table construction. Moreover, as they are completely based on linguistic knowledge, the system is robust to noise, can cover unseen units and is not faced with the data sparseness problem. Fig. 1(a) illustrates the difference between a data-driven system and a feature-table-based system.

The feature table we are using is a list of subword units tagged with their features. In manual question generation procedures, subword units are grouped according to features and combinations of features.

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004.

Accepted after full paper review

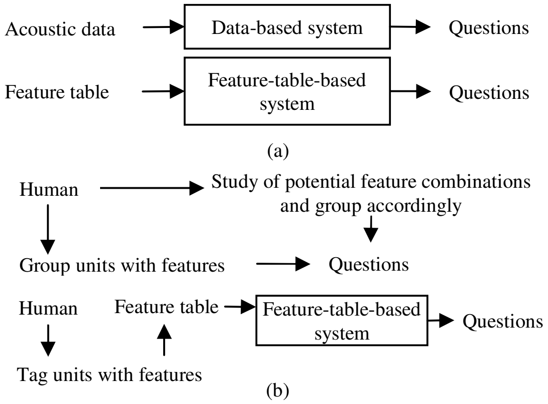

Figure 1: (a) Data-based system vs. feature-table-based system. (b) Manual procedure vs. feature-table-system.

Every possible feature combination thus has to be considered. Fig. 1(b) illustrates the difference between a manual question generation procedure and a feature-table-based system. The proposed system's purpose is analogous to the triphone clustering subsection in (Netsch and Bernard, 2004) that is to save time and human resource for a study of feature combinations. Netsch and Bernard defined two feature tables (reference and target languages) and mapped universal questions to target language questions. This limits a study of feature combination to only for universal questions, not the target language questions. In this paper, our system generates questions from a feature table and requires no universal questions. Consequently, there is no need to conduct universal questions manually and no limit of language adaptation beyond universal questions.

For the proposed system, even though there are many possible applications, only two possible applications are highlighted in this paper. First, it can automatically construct questions for a new language using only a feature table. Second, for a question selection experiment, several question sets have to be tested which is a time-consuming process for manual procedures. This paper demonstrates how questions for a new language can be set up for a question selection experiment. Thai is chosen in this paper because the available Thai corpus (Kasuriya, Sornlertlamvanich, Cotsomrong, Jitsuhiro, Kikui and Sagisaka, 2003) is relatively small and hence not suited for data-driven systems. Moreover, as presented in (Kanokphara, 2003), syllable-structure-based phonetic units are feature-rich which is a perfect example for a question selection experiment.

Systematically, this paper is organized as follows. Section 2 explains the algorithm. Section 3 introduces Thai phones and syllable-structure-based phonetic units. Section 4 describes the question selection experiment and its results while section 5 draws some conclusions.

2. Feature-table-based automatic question generation

Features can be placed on many different tiers according to their classifications, e.g., manner, place, voicing, vowel type, vowel height, etc. In English (Abu-Amer and Carson-Berndsen, 2003), for example, "p" is tagged with "stop" manner, "labial" place, "voiceless" voice, "nil" vowel type and "nil" vowel height where "nil" means there is no classification on that tier.

The strategy in this paper relies on that every feature in the feature table is well-organized and there is no identical feature on different tiers except "nil". The assumption behind the algorithm in this paper is that for every cross-tier feature combination, if there are one or more units corresponding to the combination, this combination can be used as a question. With this assumption, impossible questions are automatically discarded. For example, according to the English feature table, there will never be a "voiceless and vocalic" question because the intersection of voiceless and vocalic groups yields empty set. The algorithm is as follows.

1. Read subword units and their features.
2. List unique features from each tier.
3. Add a "blank" feature in each tier in order to bypass some tiers in some combinations. In other words, assume that every unit has a "blank" feature in every tier.
4. Group subword units according to each feature.
5. Combine each feature from different tiers with "and" and subword unit groups according to the features with set intersection.
6. Reject empty, similar-member and "nil" groups. "nil" is not a feature and there should be no "nil" or "nil" combination group.
7. Generate questions according to the groups from cross-tier feature combination.

3. Thai language and syllable-structure-based phonetic units

The sound symbols and their feature descriptions in this paper are referenced from (Luksaneeyanawin, 1993) except that "?" is replaced by "z" for the reason of program compatibility. Some phonetic symbols might be confusing because they are not in IPA format such as "q" is a vowel, not a consonant. However, this can be clarified by looking at the Manner tier as it is tagged with "vocalic". Table 1 summarizes the Thai phone feature table. Thai phones are characterized on 6 different tiers similar to the English feature table (Abu-Amer et. al, 2003). Only "aspirated" and "unaspirated" on the Manner tier are Thai specific features. In this table, there is no Static tier because diphthongs are not included since we want to describe the smallest sound

Table 1: Thai phone feature table

Phone	Voice	Stop	Manner	Place	Height	Round
@	voice	nil	vocalic	back	low	round
a	voice	nil	vocalic	central	low	unround
b	voice	stop	voiced-stop	labial	nil	nil
c	unvoice	stop	unaspirated	palatal	nil	nil
ch	unvoice	stop	aspirated	palatal	nil	nil
d	voice	stop	voiced-stop	alveolar	nil	nil
e	voice	nil	vocalic	front	middle	unround
f	unvoice	non	fricative	labial	nil	nil
h	unvoice	non	fricative	glottal	nil	nil
i	voice	nil	vocalic	front	high	unround
j	voice	non	approximation	palatal	nil	nil
k	unvoice	stop	unaspirated	velar	nil	nil
kh	unvoice	stop	aspirated	velar	nil	nil
l	voice	non	lateral	alveolar	nil	nil
m	voice	non	nasal	labial	nil	nil
n	voice	non	nasal	alveolar	nil	nil
ng	voice	non	nasal	velar	nil	nil
o	voice	nil	vocalic	back	middle	round
p	unvoice	stop	unaspirated	labial	nil	nil
ph	unvoice	stop	aspirated	labial	nil	nil
q	voice	nil	vocalic	central	middle	unround
r	voice	non	trill	alveolar	nil	nil
s	unvoice	non	fricative	alveolar	nil	nil
sil	sil	sil	sil	sil	sil	sil
t	unvoice	stop	unaspirated	alveolar	nil	nil
th	unvoice	stop	aspirated	alveolar	nil	nil
u	voice	nil	vocalic	back	high	round
v	voice	nil	vocalic	central	high	unround
w	voice	non	approximation	labial	nil	nil
x	voice	nil	vocalic	front	low	unround
z	unvoice	stop	unaspirated	glottal	nil	nil
sp	sp	sp	sp	sp	sp	sp

unit that cannot be further divided. For example, "th" is the aspirated version of "t", not the combination of two phones, "t" and "h".

Thai syllable structure is not as complex as English if tone is discarded because there are no consonant clusters in original Thai words. There are only some consonant clusters for loan words, i.e. "bl", "br", "dr", "fl", "fr", "khl", "khr", "khw", "kl", "kr", "kw", "phl", "phr", "pl", "pr", "thr", "tr", "jf^" and "ts^" which are still small compared with possible English consonant clusters. This permits Thai subword units to be designed as syllable-structure-based phonetic units. Syllable-structure-based phonetic units are units whose syllables are constrained by onset-nucleus-coda structure. In other words, there are always 3 syllable-structure-based phonetic units in each syllable and consonants are separated into onsets and codas. "^" marks coda consonant. For example, "p^" is coda version of "p". These two units are similar but not the same.

Table 2 displays the feature table of syllable-structure-based phonetic units which are just the combinations of phones and their features from Table 1. As syllable-structure-based phonetic units are feature-rich, three more tiers are added to the feature table, i.e. Position, Consonant and Length. Position contains "onset", "nucleus" and "coda" which are unit positions in a syllable. Consonant contains "consonant" and "non-con". Length indicates an unit length, "short" and "long" for single and double similar phone nucleus, and "single" and "cluster" for single and double consonants. Features for onset and coda consonants are all the same except on the Position tier. Features for all phone combination units are tagged in "A:B" format where "A" and "B" are features for left and right portions of a unit, respectively. For example, "bl" is tagged with "voiced-stop:lateral" as "b" is "voiced-stop" and "l" is "lateral" on the Manner tier. With this "A:B" format, questions about unit portions, additional to whole unit questions, can be asked such as "Is the right portion of the left context in the lateral group?", etc.

4. Question selection experiment

4.1. Experimental paradigm

The speech database used in this paper consists of 390 Thai phonetically balanced (PB) sentences. The vocabulary size is 1,476 words. The average number of words per sentence is 10. The average number of phones per word is 3.6. 42 speakers (21 males and 21 females) are separated into 34 speakers (17 males and 17 females) for training and 8 speakers (4 males and 4 females) for testing. Speakers for training are required to read 376 from 390 sentences while speakers for testing read another 14 sentences. All utterances are recorded in an office environment.

All experiments are trained and tested by using the HTK toolkit (Young, Evermann, Kershaw, Moore, Odell, Ollason, Povey, Valtchev and Woodland, 2002). The read speech utterances (16 kHz/ 16 bits) are parameterized into 12 dimensional MFCC with their energy, delta and acceleration (39 length front-end parameters).

The acoustic model topology is a 5-state left-right model with no skip state. The context-dependent system uses a cross-word network. The language model is trained from the training set using back-off bi-gram. Its perplexity is 73.68 and entropy is 6.20. The dictionary is generated automatically.

In order to observe the result explicitly, the evaluation in this paper is shown by the number of word deletion errors (D), number of word substitution errors (S), number of word insertion errors (I) and number of sentence substitution errors (SS).

Table 2: Syllable-structure-based phonetic unit feature table

Unit	Position	Consonant	Length	Voice	Stop	Manner	Place	Static	Round	Height
@	nucleus	non-con	short	voice	nil	vocalic	back	static	Round	Low
@@	nucleus	non-con	long	voice	nil	vocalic	back	static	Round	Low
a	nucleus	non-con	short	voice	nil	vocalic	central	static	Unround	Low
aa	nucleus	non-con	long	voice	nil	vocalic	central	static	Unround	Low
b	onset	consonant	single	voice	stop	voiced-stop	labial	nil	nil	Nil
bl	onset	consonant	cluster	voice	stop:non	voiced-stop:lateral	labial:alveolar	nil	nil	nil
br	onset	consonant	cluster	voice	stop:non	voiced-stop:trill	labial:alveolar	nil	nil	nil
c	onset	consonant	single	unvoice	stop	unaspirated	palatal	nil	nil	nil
ch	onset	consonant	single	unvoice	stop	aspirated	palatal	nil	nil	nil
ch^	coda	consonant	single	unvoice	stop	aspirated	palatal	nil	nil	nil
d	onset	consonant	single	voice	stop	voiced-stop	alveolar	nil	nil	nil
dr	onset	consonant	cluster	voice	stop:non	voiced-stop:trill	alveolar	nil	nil	nil
e	nucleus	non-con	short	voice	nil	vocalic	front	static	Unround	middle
ee	nucleus	non-con	long	voice	nil	vocalic	front	static	Unround	middle
f	onset	consonant	single	unvoice	non	fricative	labial	nil	nil	nil
f^	coda	consonant	single	unvoice	non	fricative	labial	nil	nil	nil
fl	onset	consonant	cluster	unvoice:voice	non	fricative:lateral	labial:alveolar	nil	nil	nil
fr	onset	consonant	cluster	unvoice:voice	non	fricative:trill	labial:alveolar	nil	nil	nil
h	onset	consonant	single	unvoice	non	fricative	glottal	nil	nil	nil
i	nucleus	non-con	short	voice	nil	vocalic	front	static	Unround	high
ia	nucleus	non-con	short	voice	nil	vocalic	front:central	non	Unround	high:low
ii	nucleus	non-con	long	voice	nil	vocalic	front	static	Unround	high
iia	nucleus	non-con	long	voice	nil	vocalic	front:central	non	Unround	high:low
j	onset	consonant	single	voice	non	approximation	palatal	nil	nil	nil
j^	coda	consonant	single	voice	non	approximation	palatal	nil	nil	nil
jf^	coda	consonant	cluster	voice:unvoice	non	approximation:fricative	palatal:labial	nil	nil	nil
k	onset	consonant	single	unvoice	stop	unaspirated	velar	nil	nil	nil
k^	coda	consonant	single	unvoice	stop	unaspirated	velar	nil	nil	nil
kh	onset	consonant	single	unvoice	stop	aspirated	velar	nil	nil	nil
khl	onset	consonant	cluster	unvoice:voice	stop:non	aspirated:lateral	velar:alveolar	nil	nil	nil
khr	onset	consonant	cluster	unvoice:voice	stop:non	aspirated:trill	velar:alveolar	nil	nil	nil
khw	onset	consonant	cluster	unvoice:voice	stop:non	aspirated:approximation	velar:labial	nil	nil	nil
kl	onset	consonant	cluster	unvoice:voice	stop:non	unaspirated:lateral	velar:alveolar	nil	nil	nil
kr	onset	consonant	cluster	unvoice:voice	stop:non	unaspirated:trill	velar:alveolar	nil	nil	nil
kw	onset	consonant	cluster	unvoice:voice	stop:non	unaspirated:approximation	velar:labial	nil	nil	nil
l	onset	consonant	single	voice	non	lateral	alveolar	nil	nil	nil
l^	coda	consonant	single	voice	non	lateral	alveolar	nil	nil	nil
m	onset	consonant	single	voice	non	nasal	labial	nil	nil	nil
m^	coda	consonant	single	voice	non	nasal	labial	nil	nil	nil
n	onset	consonant	single	voice	non	nasal	alveolar	nil	nil	nil
n^	coda	consonant	single	voice	non	nasal	alveolar	nil	nil	nil
ng	onset	consonant	single	voice	non	nasal	velar	nil	nil	nil
ng^	coda	consonant	single	voice	non	nasal	velar	nil	nil	nil
o	nucleus	non-con	short	voice	nil	vocalic	back	static	round	middle
oo	nucleus	non-con	long	voice	nil	vocalic	back	static	round	middle
p	onset	consonant	single	unvoice	stop	unaspirated	labial	nil	nil	nil
p^	coda	consonant	single	unvoice	stop	unaspirated	labial	nil	nil	nil
ph	onset	consonant	single	unvoice	stop	aspirated	labial	nil	nil	nil
phl	onset	consonant	cluster	unvoice:voice	stop:non	aspirated:lateral	labial:alveolar	nil	nil	nil
phr	onset	consonant	cluster	unvoice:voice	stop:non	aspirated:trill	labial:alveolar	nil	nil	nil
pl	onset	consonant	cluster	unvoice:voice	stop:non	unaspirated:lateral	labial:alveolar	nil	nil	nil
pr	onset	consonant	cluster	unvoice:voice	stop:non	unaspirated:trill	labial:alveolar	nil	nil	nil
q	nucleus	non-con	short	voice	nil	vocalic	central	static	unround	middle
qq	nucleus	non-con	long	voice	nil	vocalic	central	static	unround	middle
r	onset	consonant	single	voice	non	trill	alveolar	nil	nil	nil
s	onset	consonant	single	unvoice	non	fricative	alveolar	nil	nil	nil
s^	coda	consonant	single	unvoice	non	fricative	alveolar	nil	nil	nil
t	onset	consonant	single	unvoice	stop	unaspirated	alveolar	nil	nil	nil
t^	coda	consonant	single	unvoice	stop	unaspirated	alveolar	nil	nil	nil
th	onset	consonant	single	unvoice	stop	aspirated	alveolar	nil	nil	nil
thr	onset	consonant	cluster	unvoice:voice	stop:non	aspirated:trill	alveolar	nil	nil	nil
tr	onset	consonant	cluster	unvoice:voice	stop:non	unaspirated:trill	alveolar	nil	nil	nil
ts^	coda	consonant	cluster	unvoice	stop:non	unaspirated:fricative	alveolar	nil	nil	nil
u	nucleus	non-con	short	voice	nil	vocalic	back	static	round	high
uu	nucleus	non-con	long	voice	nil	vocalic	back	static	round	high
uua	nucleus	non-con	long	voice	nil	vocalic	back:central	non	round:unround	high:low
v	nucleus	non-con	short	voice	nil	vocalic	central	static	unround	high
vv	nucleus	non-con	long	voice	nil	vocalic	central	static	unround	high
vva	nucleus	non-con	long	voice	nil	vocalic	central:central	non	unround	high:low
w	onset	consonant	single	voice	non	approximation	labial	nil	nil	nil
w^	coda	consonant	single	voice	non	approximation	labial	nil	nil	nil
x	nucleus	non-con	short	voice	nil	vocalic	front	static	unround	low
xx	nucleus	non-con	long	voice	nil	vocalic	front	static	unround	low
z	onset	consonant	single	unvoice	stop	unaspirated	glottal	nil	nil	nil
z^	coda	consonant	single	unvoice	stop	unaspirated	glottal	nil	nil	nil

Table 3: Feature table level modification

	D	S	I	SS
Table 2	3	148	24	77
No Voice tier	4	148	26	77
No Position tier	4	146	25	78
No Length tier	4	146	23	75
No Stop tier	4	138	27	76
No “single” & “cluster” (A)	3	145	23	76
A+“stop:non” → “non” (B)	4	144	23	77
A+No Static tier (C)	3	144	23	76
A+C+No “voice” for nucleus	**2**	**143**	**23**	**76**

4.2. Question sets

As mentioned above, the benefit of the proposed system is that it can generate various kinds of questions with only a minor modification. Modifications in this paper can be divided into feature table level and program level.

4.2.1. Feature table level

Table 3 demonstrates the relevance of tiers or features in the Table 2 to the number of word and sentence errors.

- **Table 2**: The whole feature table from Table 2 is set as the initial baseline in this experiment.
- **No Voice tier**: The Voice tier is one of the broadest classifications as there are only two classes per tier. Our hypothesis is that Voice tier should not be included in the feature table. D increases from 3 to 4 and I increases from 24 to 26. This means that Voice tier is important for questions.
- **No Position tier**: As Position is a new tier added to the feature table, we want to examine the effect of having no Position tier. S decreases from 148 to 146. However, D increases from 3 to 4, I increases from 24 to 25 and SS increases from 77 to 78. This means that the Position tier cannot be removed from feature table.
- **No Length tier**: The experiment is conducted for the same reason as **No Position tier**. D increases from 3 to 4. Therefore, Length tier is also necessary. Conversely, as there are great decreases in S, I and SS. There will be more experiment for Length later.
- **No Stop tier**: The experiment is conducted for the same reason as **No Voice tier**. Here the tier is important tier because D and I increase when there is no Stop tier.
- **No “single” & “cluster”**: As mentioned above, a more specific experiment for Length tier is conducted. “single” and “cluster” are replaced by “nil”. This shows faultless improvement because S, I and SS reduce without increment in D. From now on, **No “single” & “cluster”** feature table is a baseline.
- **“stop:non” → “non”**: We want to prove that there is no blend between two different features; Compares with **No “single” & “cluster”** result, this supports our hypothesis.
- **No Static tier**: The experiment is conducted with the same reason as **No Voice tier**. Surprisingly, the system gives better results without Static tier. This is because on the Static tier, too many different diphthongs are grouped together with the feature “non” and therefore, this tier should be removed from feature table. Again, from now on **No Static tier** feature table is as a baseline.
- **No “voice” for nucleus**: We still suspect that the Voice tier is too broad for questions. However, as the result from **No Voice** indicates necessity of Voice tier, a deeper analysis is carried out here. In this modification, “voice” for every nucleus is replaced by “nil”. As expected, voiced consonant and voiced nucleus should not be grouped together. As this modification yields a positive result, from now on this feature table (according to the bottom row of Table 3) is considered as a baseline.

Table 4: Program level modification

	D	S	I	SS
Less than 20	4	147	26	78
Less than 25	3	149	23	78
One feature	7	147	27	81
Left & right	3	146	23	76
Portion mixed	2	146	23	76

4.2.2. Program level

Apart from the feature table, Table 4 lists the experimental results on this program level. This level provides modifications as follows:

- **Less than 20**: From feature table level, too broad tier absolutely yields unpleasant results. Since too broad tier results in a large number of units in a group, we suspect that the number of units in each group should be limited. In this modification, the number of subword units in each group is limited to be less than 20 (20 is around ¼ of the number of all units in Table 2). This does not seem to be a good modification since there are many errors increment comparing with the baseline (bottom row in Table 3).
- **Less than 25**: In this modification, the number of units is limited to be less than 25 which is larger than last experiment. Since the result does not show perfect improvement, we conclude that there

should be no limit in the number of units in each group.

- **One feature**: All "A:B" features are counted as one single feature so that there is no question for left or right portion of subword units. From Table 4, the number of errors increases and we conclude that left and right portion should be treated separately.
- **Left & right**: The intention of this modification is similar to section 3 in (Singh et. al, 1999). Left context questions should only be based on right portion of subword units and vice versa. However, the results from Table 4 do not support this idea. We, therefore, further investigate the question sets for baseline and this modification. Baseline question set contains 1,048 questions while this modification question set contains 655 questions. This modification question set is the subset of baseline question set. It is quite clear that most of the correct questions are in the smaller question set because there are only a little error reductions when 393 (1,048-655=393) more questions are added to the question set. This means that most of the correct questions are left & right specific questions (right portion group for left context and vice versa). However, since the tree-based state clustering is efficient enough to select only the correct questions, left and right context questions should be treated equally.
- **Portion mixed**: To extend more modifications about left and right portion questions such as "Does left context contain lateral portion?" In other words, two units can be in the same group if the left portion of one unit equals the right portion of another unit and vice versa. This modification also gives worse result than the baseline.

5. Conclusion

This paper shows two possible applications of feature-table-based question generation systems. Questions for a new language as Thai can be easily generated with the feature table from (Luksaneeyanawin, 1993). Even though there is no feature table existing in some languages, the implementation is still easy as shown in question selection experiment. Changing from a phone feature table to a syllable-structure-based phonetic unit feature table can be done by combining features according to phone combinations. Moreover, modifications in the feature table and at the program level are very straightforward and simple.

From the question selection experiment, there are many interesting points. First, Position, Consonant and Length are useful additions to develop questions. Second, too broad categories, such as Static, Voice for nucleus and Length for consonant, are not good for questions. Third, the number of units in each group should not be limited. Finally, left and right portion features should be treated separately. In future work, these will be tested with other languages.

Moreover, the proposed system is similar to (Neugebauer, 2004) and the two systems will be compared in the future.

6. Acknowledgements

Part of this work has been conducted by the first author at National Electronics and Computer Technology Center (NECTEC), Thailand. The first author would like to thank Prof. Yoshinori Sagisaka for his comments at NECTEC which was the original motivation for this work, and NECTEC for the use of NECTEC-ATR Thai corpus, the Science Foundation Ireland for the support under Grant No. 02/IN1/I100.

The opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

7. References

Abu-Amer T. and Carson-Berndsen J. (2003). HARTFEX: A multi-dimensional system of HMM based recognizers for articulatory feature extraction, Eurospeech.

Kanokphara, S. (2003). Syllable Structure Based Phonetic Units for Context-Dependent Continuous Thai Speech Recognition, Eurospeech, 797-800.

Kasuriya S., Sornlertlamvanich V., Cotsomrong P., Jitsuhiro T., Kikui G. and Sagisaka Y. (2003). NECTEC-ATR Thai Speech Corpus, COCOSDA.

Luksaneeyanawin, S. (1993). Speech Computing and Speech Technology in Thailand, Proc. SNLP, 276-321.

Netsch, L. and Bernard, A. (2004). Automatic and Language Independent Triphone Training using Phonetic Tables, ICASSP.

Neugebauer, M. (2004). Akustische Modellierung mit phonetischen Entscheidungsbaeumen auf der Grundlage phonologischer Typhierarchien. In Proc. Konvens, to appear (in German).

Odell, J. J. (1995). *The Use of Context in Large Vocabulary Speech Recognition*, Ph.D. Thesis, Cambridge University, Cambridge.

Singh, R., Raj, B., and Stern, R. M. (1999). Automatic Clustering and Generation of Contextual Questions for Tied States in Hidden Markov Models, ICSLP.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P. (2002). *The HTK Book (for HTK Version 3.2),* Microsoft Corporation and Cambridge University Engineering Department, England.