PHONETIC MODELLING IN THE PHILIPS CHINESE CONTINUOUS-SPEECH RECOGNITION SYSTEM

Frank Seide, Nick J.C. Wang

Philips Innovation Center, Taipei
24FA, 66, Sec. 1, Chung Hsiao W. Rd., Taipei, Taiwan, Republic of China
E-mail: {seide,nick}@prlt.research.philips.com

ABSTRACT

We have extended the Philips large-vocabulary continuous-speech recognition system towards Chinese. On the way from our existing Western-language technology to Mandarin, the first step was to build a suitable phonetic model. This paper describes the development of our phonetic model (excluding tones) for Mandarin Chinese.

We will present a systematic comparison of three forms of sub-syllabic units for Chinese, *phonemes*, *initials/finals*, and a non-tonal form of *preme/toneme models*, as well as whole-syllable models for reference. We include experiments on bottom-up and decision-tree based top-down state clustering and modelling of cross-syllable contexts.

All forms of sub-syllabic units are represented in the Philips Mandarin phone set "SAMPA-C." SAMPA-C is based on the European SAMPA standard and introduced in this paper.

Our studies show that traditional half-syllable approaches slightly outperform Western-style triphones. Modelling of right-context dependency gives greater improvement than left-context dependency, and cross-syllable modelling yields a 4-5% performance gain. In a free syllable decoding task, we achieve 39% syllable error rate for telephone speech and 24% for microphone dictations.

1. INTRODUCTION

The Philips large-vocabulary continuous-speech recognition technology has successfully been applied to many Western languages including U.S. English [1], German and French [2], and Dutch [3] for various applications including dictation and dialogue systems [4]. In the process of localizing our system to Mandarin Chinese, we have been facing a few fundamental differences compared to the Western languages.

The obvious differences are that Chinese is a tonal language and uses an ideographic writing system. However, the fundament of any successful large-vocabulary recognizer is its phonetic model. I.e., we need to find a set of sub-word units that provides the best balance between accuracy and optimal use of training material. One may hope that the strongly syllable-oriented structure of Chinese gives room for additional improvements.

In this paper, we compare three basic approaches of sub-word – or more precisely *sub-syllabic* – modelling for Chinese: Western-style *phonemes*, traditional *initials/finals* [5, 6], and a non-tonal version of *premes/tonemes* [7]. In addition, we compare whole syllable models. Performance is evaluated for the task of unconstrained syllable recognition on continuously read sentences over the telephone. They are investigated w.r.t. phoneme duration estimation, base phone set performance, data-driven state clustering, context-dependency ranking, and cross-syllable modelling. The results are verified on a microphone dictation corpus.

Based on the European SAMPA standards for phonetic transcriptions, we have defined a phone set for Mandarin Chinese. This set, which we named SAMPA-C, is suited to represent the three considered sub-syllabic unit types.

This paper is organized as follows. Section 2 will introduce the SAMPA-C phone set. In section 3, we will review the three sub-syllabic modelling approaches. Section 4 will describe our experimental setup, and section 5 the results. Finally, section 6 will summarize the paper.

2. THE SAMPA-C PHONE SET

When designing the Mandarin phone set, we set a number of requirements. First, the context-independent representation of pronunciation should be accurate, i.e. it should not be necessary to assume a particular form of modelling of context dependencies. Next, it should be flexible to represent different types of sub-syllabic unit selection, which excludes half-syllable based systems. Finally it had to be as similar as possible to our Western-language phone sets, allowing for multi-lingual phone sets.

For the Western languages, Philips makes consequent use of the European SAMPA standard [8], developed by the European-Union funded "Speech Assessment Methodologies" project (SAM). SAMPA provides machine-readable, 7-bit ASCII compatible representations of the International Phonetic Alphabet (IPA) for many Western languages. An extension named X-SAMPA defines mappings for the entire IPA [9].

Thus, we decided to base our phone set definition on IPA [10, 11], translated it to X-SAMPA, and made the following modifications to fit our additional requirements:

- In a few cases, pronunciation of an IPA symbol for Chinese varies beyond coarticulation. In these cases, we use the following distinct position- or context-dependent symbols (some of which are undefined in X-SAMPA):
  - The symbol `r` is used only at final position. At initial position, /r/ is pronounced slightly postalveolarized, denoted by the new symbol `R` instead.
  - `u` is used only for glide. At initial position, we replace it by labial-velar `w`.
  - `i` also only in glide position. For the long vowel in final position, we add a colon (`i:`), and if followed by a nasal, `I` is more accurate.
  - The empty final is represented as the long open-mid central `3:` after `S`, `Z`, or `R`, and as a prolonged lax `I:` (new symbol) after `s` or `z`.
  - Also `n` only at initial position due to weak articulation and retroflexization at final position. Here, we introduced the symbol `M`.
- We removed the aspiration marker from `p_h` and `t_h` (redundancy) and `s_h` and `s\_h` (little acoustic difference).
- To avoid backslash and quote characters (unhandy with popular text processing tools), retroflex `s`` and `z`` were approximated by postalveolar `S` and `Z`, respectively, and `s\` and `z\` by palatal `C`.
- As common in European SAMPA (and IPA) we distinguish long vowels from their short counterparts by appending a colon.

Table 1 shows all SAMPA-C symbols with an example as well as the original X-SAMPA representation. For comparison, the table also shows SAMPA-T, another recent proposal developed to transcribe the Chinese dialects spoken on Taiwan, including Taiwanese and Hakka [12].

For the design of a phone set, a major issue is the compatibility with existing language resources. The most popular transcription systems for Chinese are hànyǔ-pīnyīn (official standard Mainland China) and zhùyīn ("bopomofo", popular in Taiwan, also used for teaching). Both are not suitable as phone sets due to their (slight) incompatibility with the "initial/final" representation (see below), and the latter requires unhandy double-byte encoding. However, a one-to-one mapping to SAMPA-C is possible on syllable level.

3. SUB-SYLLABIC MODELLING

Chinese is often referred to as a mono-syllabic language. In fact, Mandarin is based on less than 410 different syllables[1].

Traditionally, a syllable is split into two halves, an *initial* consonant part and a vocalic/semi-vocalic *final* part. There are 22 initials and 38 finals, including the special cases of a *null initial* and the *empty final*, as listed in table 2. Highly accurate Chinese speech-recognition systems have been built

Table 1: The SAMPA-C phone set.

SAMPA			Example		
-C	X-	-T	SAM-PA-C	Pinyin	Word
b	b	b	ba:	八 bā	eight
p	p_h	p	pi:	皮 pí	skin
m	m	m	m@M	門 mén	gate/door
f	f	f	fVN	風 fēng	wind
d	d	d	da:	大 dà	big
t	t_h	t	tai	台 tái	platform
n	n	n	naM	南 nán	south
l	l	l	lu:	路 lù	road
z	z	z	dzI:	子 zǐ	son
s	s s_h	s s	saM tsu@M	三 sān 村 cūn	three village
R	r	Z`	REM	人 rén	person
Z	z`	z`	dZUN	中 zhōng	middle
S	s` s`_h	s` s`	Suei tSa:	水 shǔi 茶 chá	water tea
C	s\ s_h z\	s\ s\ z\	Ciao tCi: dCIN	小 xiǎo 七 qī 京 jīng	small seven capital
g	g	g	guO	國 gúo	country
k	k	k	kou	口 kǒu	opening
h	h	h	hV:	河 hé	river
?	-	-	?aM	安 ān	peace
w	u	u	waN	王 wáng	king
j	j	i	jou	有 yǒu	to have
a:	a	a	da:	大 dà	big
V:	V	@	dV:	德 dé	virtue
3:	3	U`	S3:	士 shì	scholar
I:	1	U	sI:	四 sì	four
i:	i	i	?i:	一 yī	one
O:	O	O	bO:	伯 bó	uncle
u:	u	u	?u:	五 wǔ	five
y:	y	y	tCy:	去 qù	to go
a	a	a	SaM	山 shān	mountain
@	@	@	m@M d@	門 mén 的 dė	gate/door of
V	V	@	fVN	風 fēng	wind
E	E	E	CiE	謝 xìe	thanks
I	i	i	CIM	心 xīn	heart
i	i	i	Cia:	下 xìa	under
U	U	u	gUN	工 gōng	work
u	u	u	tSuaM	川 chuān	river
y	y	y	?yEM	元 yuán	money unit
ai	ai	ai	sai	賽 sài	contest
aU	aU	au	dzaU	早 zǎo	morning
ei	ei	ei	bei	北 běi	north
ou	ou	ou	kou	口 kǒu	opening
uO	uO	uo	guO	國 gúo	country
M	n	n	tiEM	天 tiān	heaven
N	N	N	SaN	上 shàng	above
r	r	(@`)	?r	二 èr	two

[1] The exact number varies slightly depending on incorporation of dialect-specific syllables and coverage of spontaneous speech. Our dictionary has 398 syllables

on *initial/final models*, using directly the unit set from table 2. Often, right-context dependent initial models are used, as in Golden Mandarin III [5] or the JANET system [6].

In [7], a modified scheme was suggested: Syllables containing a "glide" (i.e. the final begins with a short medial i, u, or y) are split *after* the glide; the set of finals is thus reduced to the core finals without glide (so-called *tonemes* – we will call them *core finals* since our models currently do not model tone), while the set of initials is extended accordingly (so-called *premes*). Table 3 shows the additional premes with glide. Although this approach is reported to be particularly beneficial in combination with tone modelling, we have included it into our investigations.

Table 2: Initials and finals.

Initials	b, p, m, f, d, t, n, l, z, c, s, zh, ch, sh, r, j, q, x, g, k, h, *null initial*
Core finals (no glide)	a, e, i, o, u, ü, er, *empty final*, ai, ei, ao, ou, an, en, ang, eng, ong
Finals with glide	ia, iao, ie, i(o)u[2], iai, ian, in, iang, ing, iong, ua, uo, uai, u(e)i, uan, u(e)n, uang, ueng, üe, üan, ü(e)n

Table 3: Premes with glide.

Premes with glide	bi, pi, mi, di, du, ti, tu, ni, nu, nü, li, lu, lü, zu, cu, su, zhu, chu, shu, ru, ji, ju, qi, qu, xi, xu, gu, ku, hu, yi, yu, wu

The third possibility is to use *phonemes* similar to phonemes of Western languages. A potential benefit from using smaller units is the higher degree of model sharing and finer control in the context of model clustering.

In our half-syllable unit inventories, there is a modification w.r.t. the null initial: In syllables beginning with the glides i or u, the glides become initials, denoted by the SAMPA-C symbols j and w, respectively. In addition, we introduced right-context dependent versions of the glottal stop, mainly for cross-word modelling in order to extend the right-context across the glottal stop into the following vowel. In total we have 30 initials and 40 finals, 61 premes and 25 core finals, and the SAMPA-C phone set consists of 50 units.

4. EXPERIMENTAL SETUP

The Philips system is a HMM-based large-vocabulary continuous-speech recognizer. We use standard MFCC features with first-order derivatives, sentence-based cepstral mean subtraction (CMS) for simple channel normalization, and Gaussian mixture densities with density-specific diagonal covariance matrices.

We evaluated the different acoustic models for the task of free decoding of continuously spoken syllables (read text), using a vocabulary of 398 syllables and a hand-tuned word-insertion penalty serving as a zerogram language model. The error rates given are syllable error rates (SER), i.e. we do not try to identify tone or character. All models are speaker and gender independent without speaker or acoustic adaptation.

The main investigations were done on the telephone corpus *MAT* (Mandarin Across Taiwan) [13]. 90% of MAT subcorpora 4 (isolated phrases) and 5 (continuous read sentences) were used for training (12.6h netto speech from 721 Taiwanese speakers plus 6.5h of silence). A disjunct 72-speaker 22-minutes subset of subcorpus 5 was set aside for evaluation. The data is sampled at 8 kHz and has an average SNR of 34 dB.

We verified the results on a high-quality continuous-dictation database. The 22 kHz recordings were made in a quiet room (SNR 44 dB) and consist of newspaper sentences spoken by Taiwanese speakers (54.7h netto speech, 17.2h silence). Table 4 summarizes the corpus.

Table 4: Corpus characteristics.

	Telephone		Microphone	
	Training	Test	Training	Test
#Speakers	721	26	241	20
#Utterances	28896	259	27606	200
#Syl./Utt.	5.6	14.2	30.1	35.5
#Perplexity	—	398	—	398

All error rates shown are "best possible" results. I.e. instead of fixing the number of mixture components or the total number of model parameters, we optimized the error rate for each setup over a range of resolutions (32 to 256 mixture components), and report the individual optimum values.

5. RESULTS

5.1. Phoneme Length

In the Philips system, a sub-word unit p consists of N_p so-called *segments* of two states. The two states forming a segment share the same emission distribution. Allowed state transitions are *loop*, *next*, and *jump* (skip one state). The time-distortion penalties (which correspond to the log state-transition probabilities) are pooled over all states and fixed; a model's duration is reflected in its number of segments.

For Western languages, we usually chose the same length of $N_p = 3$ segments for all phonemes. For Chinese, half-syllable models may be significantly longer, and model lengths used in other systems (e.g. 3 states for the initial and 4 for the final in Golden Mandarin III [5]) are not applicable for our segment approach. Instead, we used the following simple estimation procedure to determine the model length:

1. Train context-independent 3-segment models.
2. For each model determine the duration (number of segments) that minimizes the overall time-distortion penalty given the length distribution actually observed in the 3-segment training.
3. Clone the duration to all context-dependent models of the same family.

[2] The vowels shown in parentheses are pronounced but omitted in the standard written pinyin form.

We tested the accuracy of this procedure on SAMPA-C triphones (table 5). We observed a slight relative gain of 2.6% over the standard 3-segment duration. The average SAMPA-C phoneme duration was determined as 4.3 segments (8.6 states; the average observed length was 95.3 miliseconds). In all experiments below, we applied this duration estimation scheme to all different forms of sub-syllabic units.

Table 5: Results on phoneme length estimation.

#States/phone	3	est'd
#States	1379	1883
#Densities	139k	108k
SER	45.0%	43.8%

5.2. Basic Phone Set

We compared the three different sets of sub-syllabic units. Tables 6, 7, and 8 show the results for different degrees of (within-syllable) context-dependency for triphones ("SAMPA-C"), initial/final ("IF") models, and preme/core-final ("PCF") models, respectively. Table 9 shows the result for a whole-syllable model, which provides maximum context dependency. The following dependencies were investigated:

- *none*: Only context-independent units.
- *left/right*: Single-sided left or right context dependency (diphones). For within-syllable IF and PCF models, either initials or finals are context independent.
- *both*: The first half syllable is right, and the second half is left-context dependent.
- *triph.*: Triphones.
- *full*: SAMPA-C phones with full-syllable context dependency (fallback for rare/unseen units).
- *trans.*: The syllable consists of three sub-syllabic units: a context-independent initial, a context-dependent transition (around one third of each syllable's states), and a context-independent final.

To reduce the effect of overadaptation of rarely seen units, we applied data-driven bottom-up state clustering where applicable (see also the next section). For the transitional initial/final model, no clustering was applied, and transitions occuring less than 200 times are backed off to a context-independent initial and final half.

We observe the following. First, SAMPA-C monophones cover the shortest context and have thus higher error rate than context-independent IF and PCF models. Second, IF models are slightly outperformed by PCF models, in part due to the slightly higher number of model states (480 vs. 401), but we also believe that the coarticulation effect from glides on initials is higher than their impact on the finals.

Third, modelling dependencies on right context seems to be more important than on left context. SAMPA-C right diphones outperform left diphones by 4.1%, and for PCF we observe a relative difference of 7.7%.

Table 6: Results for SAMPA-C models.

Context	none	left	right	triph.	full
#States	220	1019	884	1883	3485
#Densities	54k	174k	156k	108k	192k
SER	54.5%	47.6%	45.6%	43.8%	41.3%

Table 7: Results for IF models.

Context	none	right	both	trans.
#States	401	1591	3014	1768
#Densities	95k	157k	167k	136k
SER	49.5%	42.1%	41.4%	43.4%

Table 8: Results for PCF models.

Context	none	left	right	both
#States	480	1488	1629	3457
#Densities	105k	143k	237k	165k
SER	47.3%	44.2%	40.8%	40.6%

Table 9: Result for whole-syllable models.

#States	5275
#Densities	228k
SER	40.7%

Finally, with increasing context dependency, error rates of all models converge to the whole-syllable model, which is reached by PCF, and missed by merely 2% rel. by IF and SAMPA-C phones. The transitional model cannot reach the performance, the context dependency is obviously not sufficiently modelled by the transitional states alone.

5.3. State Clustering

In the previous section, a bottom-up state clustering scheme was applied per default. What is the gain from clustering?

To obtain good model parameter estimates, an appropriate amount of training material is needed. Models trained on too little data adapt too closely to the training set, and their ability to represent the expected testing observations is poor.

In an untied system, we would *fall back* models with too few observations to a shorter-context model. *Tying* means to share parameters among similar states, allowing to increase coverage of context-dependent models. We use the algorithm described in [14] in the implementation of [15]. The generated clusters match very well our expectations, an example for the right-context dependent final `E` is shown in table 10.

Table 11 shows the results. By tying, within-syllable coverage is raised to nearly 100%, but the error-rate gain is limited, around 4% relative. For comparison, [15] reports a 7.3% gain in the U.S. English 1993 Wall-Street Journal benchmarking, at an average cluster size of 4.4. In our setup, the optimal cluster size (average states per cluster) was below 2.

5.4. Context-Dependency Ranking

Which coarticulation effect has the greatest impact on recognizer accuracy? Right or left, within or across syllable

Table 10: Example clusters for the last segment of the right-context dependent final family E.

Right context	Explanation
t, k, d	non-labial plosives
gu, g, p, b	voiced plosives
tu, dz	(plosive + fricative, glide u tends to fricative)
dCy, dZ, dC, dCi, tC	voiced plosive + sh-sound
C, Ci, S	sh-sounds without plosive
tS, ti, ts	unvoiced plosive + fricative (glide i following t seems to have a slight tendency towards a fricative)
s, f	s-like fricative (acoustic similarity of s and f on band limited telephone speech)
n, m	nasals in initial position
l, R	approximant
j, ?(y), ?(i)	i/y at syllable beginning
hu, w, ?(u)	u in glide position (the transition center phone E → u tends towards w)
?(o), ?(a), h, pause	glottal stop similar to silence (the transitions E → o and E → a seem to be similar to a glottal stop)
?(r), ?(e)	e and retroflex e

boundary? We investigated this questions with PCF models. For both half-syllable types, we evaluated both possible single-sided context dependencies independently (table 12) Also across syllable boundaries, the right context is more important, but cross-syllable context alone leads to rel. 14.5% higher error rate than the within-syllable context alone.

5.5. Cross-syllable modelling

When modelling cross-syllable context dependency, the number of models increases significantly, so tying becomes mandatory.[3] Furthermore, since words may be more or less freely combined, it is not guaranteed that every cross-syllable context has actually been observed at all during training.

The standard solution to this problem are *decision trees* (classification and regression trees, "CART"). It is a top-down clustering method, in which a set of *phonetic questions* defines candidates for possible clusters. Clusters are formed by successive splitting a cluster according to the question set into those sets of subclusters that yield maximum training-set likelihood. By following this hierarchy (tree) of clusters, replacements for unseen models can be found [16, 17, 18].

Normally, the question set is created by a phonetic expert. However, [19] describes an automatic procedure that reaches if not exceeds the performance of hand-crafted questions. It requires a bottom-up tying of single-sided context-dependent units, but with unlimited cluster size, such that at the end, only one large context-independent cluster is left for each state. The question set is a pruned subset of the intermediate clusters observed during the clustering.

Table 11: Effect of tying.

Model & context	PCF right	PCF both	SAMPA-C both
untied			
#States	1245	2384	1559
#Densities	139k	143k	96k
Coverage	87.2%	79.8%	82.0%
SER	42.5%	42.4%	45.4%
tied			
#States	1629	3457	1883
#Densities	237k	165k	108k
Coverage	99.9%	99.9%	99.6%
SER	40.8%	40.6%	43.8%
SER Gain	4.1%	4.2%	3.5%

Table 12: Context ranking (PCF).

Context	right within	left within	right across	left across
#States	1629	1488	3109	2888
#Dens	237k	143k	163k	222k
SER	40.8%	44.2%	46.7%	47.2%

Table 13: Cross-syllable modelling.

Model & context	SAMPA-C triph. within	across	across
Tying	BUT	BUT	CART
#States	1883	3829	3915
#Densities	108k	116k	209k
Coverage	99.6%	90.1%	100%
SER	43.8%	41.7%	41.8%
Model & context	PCF right within	across	across
Tying	BUT	BUT	CART
#States	1629	4245	3953
#Densities	237k	223k	319k
Coverage	99.9%	98.5%	100%
SER	40.8%	39.3%	39.0%

We conducted experiments on PCF and on SAMPA-C phone models. Table 13 shows the result. Cross-syllable modelling yields slight but consistent improvements (SAMPA-C: 4.8%; PCF: 4.3%). For PCF, bottom-up tying ("BUT") already yields a coverage above 98%, so we did not expect substantial gains from CART. But we were suprised to find no improvement for SAMPA-C triphones, where CART helps to increase test-set triphone coverage from 91% to 100%.

For comparison, [18] reports similarly small improvements (2% to 6%) from cross-word modelling in combination with bottom-up tying on the U.S. English WSJ 5K and NAB'94

[3] In control experiments on cross-syllable models without tying, error rates comparable to those in table 13 were around 45%, i.e. we lost accuracy due to poor model coverage and strong fall-back.

benchmarking tasks, but consistent improvements of over 10% when using decision tree-based tying. We attribute our poor decision-tree performance to the question set – we did not use pruning in question-set generation, nor have we compared it with a carefully designed hand-crafted question set.

5.6. Microphone Dictation Corpus

We verified the results by repeating the experiments from table 13 on the high-quality microphone dictation corpus described in table 4. For each test, we used the optimal model resolution from the corresponding telephone-corpus result.

Table 14: Results on the microphone corpus.

Model &	SAMPA-C triph.		
context	within	across	across
Tying	BUT	BUT	CART
#States	1830	3892	3470
#Densities	112k	123k	209k
Coverage	99.8%	98.4%	100%
SER	28.0%	26.8%	26.7%

Model &	PCF right		
context	within	across	across
Tying	BUT	BUT	CART
#States	1565	3679	3762
#Densities	178k	223k	236k
Coverage	99.9%	99.8%	100%
SER	25.0%	24.4%	23.5%

Table 14 shows the results. We observe the typical factor of 1.5 to 2 between the error rates for telephone and microphone speech. Aside from that, the relative factors between the error rates are similar to the telephone results, despite of the difference in size of the training corpus, which is over 4 times larger and leads to model coverage of over 98% in all cases.

6. CONCLUSIONS AND OUTLOOK

We have presented a comparison between different types of sub-syllabic units for continuous-speech recognition of Mandarin Chinese, including *phonemes*, *initials/finals*, and *preme/core-final models*. All experiments have been based on SAMPA-C, the Philips Mandarin phone set.

The traditional half-syllable approaches have shown to outperform Western-style triphones slightly. Modelling right-context dependency is more important than the left-context, within and across syllable boundaries. Improvements of 4% have been achieved by data-driven state tying, and another 4-5% by taking cross-syllable context into account.

The results first obtained on a telephone corpus have generalized well to a microphone dictation corpus. In the free syllable decoding task, we have achieved 39% syllable error rate for telephone speech and 24% for microphone dictations.

After all, it has been a fascinating task to approach Mandarin speech recognition from Western-language LVCSR technology. Our efforts have layed the ground for Mandarin system development. We will now address the other – more Chinese-specific – issues of tone and language modelling.

7. REFERENCES

1. X. Aubert et al. Large vocabulary continuous speech recognition of Wall Street Journal data. In *Proc. ICASSP94*, Vol. II, pp. 129–132, Adelaide, 1994.
2. Ch. Dugast et al. The Philips large-vocabulary recognition system for American English, French and German. In *Proc. EUROSPEECH*, pp. 197–200, Madrid, 1995.
3. H. Strik et al. Localizing an automatic inquiry system for public transport information. In *Proc. ICASSP96*, Vol. 2, pp. 853–856, Philadelphia, 1996
4. A. Kellner, B. Rueber, and F. Seide. A voice-controlled automatic telephone switchboard and directory information system. In *Proc. IVTTA96*, Basking Ridge, 1996.
5. H.M. Wang et al. Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary but limited training data. In *Proc. ICASSP95*, pp. 61–64, Detroit, 1995.
6. Z.Y. Wang et al. Methods towards the very large vocabulary Chinese speech recognition. In *Proc. EUROSPEECH*, pp. 215–216, Madrid, 1995.
7. C.J. Chen et al. New methods in continuous Mandarin speech recognition. In *Proc. EUROSPEECH*, pp. 1543–1546, Rhodos, 1997.
8. J. Wells. EAGLES Handbook on Spoken Language Systems (DRAFT) – SAMPA computer readable phonetic alphabet. At http://www.phon.ucl.ac.uk/home/sampa/ home.htm, 1997.
9. J. Wells. Computer Coding the IPA: a proposed extension of SAMPA. At ftp://pitch.phon.ucl.ac.uk:/pub/ sam/ipasam-x.ps, 1995.
10. R.H. Mathews. Mathews' Chinese-English Dictionary. Caves, 13th printing, 1975.
11. N.N. Loh-John. Langenscheidts praktisches Lehrbuch Chinesisch. Berlin, 1995.
12. C.Y. Tseng and F.C. Chou. Machine readable phonetic transcription system for Chinese dialects spoken in Taiwan. In *Proc. EALREW*, Tsukuba, Japan, 1998.
13. H.-C. Wang. MAT – a project to collect Mandarin speech data through telephone networks in Taiwan MAT database documentation.
14. S.J. Young et al. The use of state tying in continuous speech recognition. In *Proc. EUROSPEECH*, Vol. 3, pp. 2203–2206, Berlin, 1993.
15. Ch. Dugast et al. Application of clustering techniques to mixture density modelling for continuous-speech recognition. In *Proc. ICASSP95*, pp. 524–527, Detroit, 1995.
16. H.-W. Hon. Vocabulary-independent speech recognition: The VOCIND system. Ph.D. thesis, School of Computer Science, Carnegie Mellon University, Pittsburg, 1992.
17. S.J. Young et al. Tree-based state tying for highly accurate acoustic modelling. In *Proc. ARPA Human Language Technology Workshop*, pp. 405–410, Plainsboro NJ, 1994.
18. P. Beyerlein et al. Modelling and decoding of crossword context dependent phones in the Philips large vocabulary continuous speech recognition system. In *Proc. EUROSPEECH*, pp. 1163–1166, Rhodos, 1997.
19. K. Beulen and H. Ney. Automatic Question Generation for Decision Tree Based State Tying. In *Proc. ICASSP98*, Vol. 2, pp. 805–808, Seattle, 1998.