SOME DESIGN CRITERIA IN SEGMENTING AND LABELLING A DATABASE OF SPOKEN CANTONESE

Jonathan Harrington and Lydia K.H. So†

Speech Hearing and Language Research Centre, Macquarie University, Sydney, Australia

†Department of Speech and Hearing Sciences, University of Hong Kong, Hong Kong

ABSTRACT - The paper describes a collaborative project for the creation of a segmented and labelled database of spoken Cantonese. Details are provided both of the phonetic and phonological structure of Cantonese and the way in which the mu+ system for speech corpus analysis has been adapted to extract and analyse data from a tone language.

INTRODUCTION

In an earlier paper (So & Thelwall, 1992), the motivation for setting up a database of Cantonese was described. Essentially, the main needs were in speech therapy practice, but also in teaching English as a foreign language. Relatedly, although the phonology of Cantonese has been thoroughly investigated (e.g. Cheung, 1986), there is a paucity of acoustic and experimental phonetics studies of Cantonese. Accordingly, the project for a database of spoken Hong Kong Cantonese has been set up to provide laboratories in Hong Kong with access to centrally stored and managed speech data.

As described in So & Thelwall (1992), the speech data is to include both acoustic and multi-channelled articulatory data and must be annotated at multiple hierarchical levels. The inclusion of articulatory channels (e.g. electropalatographic and laryngographic traces) is important for providing normative data against which different kinds of speech disorders can be assessed. The motivation for hierarchical labelling is to allow investigations of e.g. the interaction between phonetic segments and tonal structure, and more generally, to predict the major influences of surface timing, which would be relevant in the context of developing a text-to-speech system for Cantonese.

Recently, a system has been described both for automatically assisted hierarchical labelling, and for extracting the labelled data and associated signal files from a speech database (Harrington, Cassidy, Fletcher and McVeigh, 1993). In this system, known as *mu+*, an orthographic string (entered at the keyboard) is matched against the string of phonetic symbols that is the result of segmenting and labelling the speech data. This matching process includes a set of modules for (i) mapping phonetic onto phonemic symbols; (ii) parsing the impoverished phonemic string into words (the phonemic string is impoverished in the sense that it will include many reductions and assimilations due to continuous speech processes); (iii) building a foot and syllable structure between the word and phoneme level, partly by retrieving prestored information of lexical stress patterns from an on-line lexicon. Once a hierarchical structure has been built for each utterance in this way, any signal files (either sampled speech data or files, such as formant frequencies, derived from the sampled speech data) can be extracted from the database with respect to the levels and labels of the hierarchical structure (e.g. "retrieve the formant values at the segment midpoint for closed stressed syllables in two-syllable content words").

As a result of preliminary collaborative research in mid-1994, mu+ has been adapted to retrieving speech labels from hierarchical structures for Cantonese speech data. This has necessarily entailed encoding certain aspects of Cantonese phonology in the system. The remaining sections give some details of the rules for mapping Cantonese phonetic segments onto their corresponding phonemes, as well the tree structure which is proposed for each word in the database. The computational requirements for adapting mu+ from the earlier 1993 version are also reported, and some details are given of the scope and recording specifications for the existing Cantonese database.

SPEECH DATA RECORDING

Since there is a paucity of available Cantonese speech data, the first stage of the project has focused on recording citation-form data from isolated words; at a later stage of the project, continuous speech

data will also be included. The speech material was made up by selecting all real words from the theoretically possible combinations of consonantal onsets, vowels, final consonants, and tones. The number of real words from this set of combinations was 1863.

The 1863 target words were randomised and produced in the carrier phrase: "I want to read _ for you to hear". All recordings were done in a sound-proof room in the Dept. of Speech and Hearing Sciences at the University of Hong Kong. The subjects were informed about the procedure and had a chance to familarize themselves with the carrier sentence and the target words. Recordings for each subject were divided into two to three sections to avoid fatigue. Currently, recordings have been made from 8 male and 8 female subjects.

SEGMENTATION AND LABELLING

The speech data was digitised at 20 kHz on a SUN SPARC system running Entropic's WAVES+ which was also used to track automatically the first four formant centre frequencies and their bandwidths, and to calculate automatically a fundamental frequency contour and RMS data. Segmentation and labelling were carried out in WAVES+ from a combination of the speech waveform, spectrogram with superimposed formant frequencies, and fundamental frequency contour. The Machine Readable Phonetic Alphabet (MRPA) for Cantonese, and its relationship to the International Phonetic Alphabet is shown in Tables 1 and 2 for consonants and vowels respectively; Table 3 has details of the MRPA for the 9 tones in Cantonese. Only phonetic segmentation is carried out at this stage: the tone is entered at a subsequent stage on a partially constructed tree for each word.

DERIVATION OF THE TREE-STRUCTURE

For each utterance in the database, a tree-structure is created which consists of a number of different levels between the Phonetic and Orthographic levels of representation. The Orthographic representation is essentially a romanization (Pin Yin) representation of the Chinese characters; in future versions, this will be replaced by the Chinese characters themselves.

The Phoneme level is created semi-automatically, as is the bracketing in the Tone level which defines the start and end of a syllable with respect to phonemes. In order to create these levels, the user enters the orthographic form of the text (in Pin Yin notation) from which the citation-form phonemic pronuncation is accessed by dictionary look-up. The citation-form phonemic representation must then be matched to the phonetic string which is the result of segmentation and labelling; in fact, the phonetic string is first converted by rule to a corresponding (reduced form) phonemic string, and this is then matched to the citation-form string to find the boundaries between words using a dynamic programming algorithm. The phonemes are automatically parsed into syllables (defined at the Tone level in Fig. 1) which, given the highly restrictive syllable structure of Cantonese, is a relatively straightforward operation.

Once the hierarchical structure has been created in this way, the user enters manually the tone of each syllable at the tone level. In a future version, the tonal information will be automatically retrieved from an on-line lexicon which includes details of each character's tone. The next section gives details of some of the phonological rules which are used in creating the tree-structure.

Phonetic-to-Phoneme rules

As described in Cheung (1986), Cantonese has 19 consonant phonemes and eight vowel phonemes. These are as follows (see Tables 1 & 2 for the relationship to IPA notation):

IPA	MRPA (Phonetic)	MRPA (Phonemic)
p^h	p + H	p
p	p	b
t^h	t + H	t
t	t	d
k^h	k + H	k
k	k	k
m	m	m
n	n	n
ŋ	ng	ng
f	f	f
s	s	s
h	h	h
w	w	w
j	j	j
l	l	l
ts^h	ts + H	c
j	ts	z
k^hw	kw + H	kw
kw	kw	gw

Table 1: MRPA for Cantonese: Consonants

Vowels		Diphthongs		
IPA	MRPA (Phonetic)	IPA	MRPA (Phonetic)	MRPA (Phonemic)
i	i	ai	aai	aaj
y	y	ɐi	ai	aj
u	u	ui	ui	uj
ɛ	e	ei	ei	ej
a	aa	ɔi	oi	oj
ɐ	a	ou	ou	Ow
ɔ	o	au	aau	aaw
œ	oe	ɐu	au	aw
I	I	iu	iu	iw
ʊ	U	œy	oey	oej
e	E			
o	O			
ø	OE			

Table 2: MRPA for Cantonese: Vowels and Diphthongs

		MRPA
1	High Level	HH
2	High Rise	MH
3	Mid Level	MM
4	Low Fall	ML
5	Low Rise	LM
6	Low Level	LL
7	High Entering	H
8	Mid Entering	M
9	Low Entering	L

Table 3: MRPA for Cantonese: Tones

	unasp. stop	*asp. stop*	*fricative*	*nasal*	*approx.*
labial	b	p	f	m	
labiovelar	gw	kw			w
dental	d	t			l
alveolar	z	c	s	n	j
velar	g	k	h	ng	

Table 4: consonant phonemes in Cantonese

Within the consonant set, there is a basic phonemic distinction between aspirated and unaspirated oral stops. The labiovelars are realised phonetically as labial stops with superimposed secondary velarisation (Cheung, 1986:111) and they can be either aspirated or unaspirated. The alveolar stops are realised as fricated stops which can also be either aspirated or unaspirated. The main phonetic-to-phoneme rules that are intrinsic to the derivation of the tree-structure in Fig. 1 concern parsing phonetic sequences of a closure followed by aspiration e.g. [p H] into the corresponding voiceless stop phoneme (i.e. /p/ in this case).

	UNROUNDED		*ROUNDED*	
	Front	*Back*	*Front*	*Back*
Close	i		y	u
Mid	e		oe	o
Open	aa, a			

Table 5: vowel phonemes in Cantonese

Seven of the eight vowel phonemesin Cantonese are realised phonetically as long vowels; /a/ is realised as a short vowel and contrasts with /aa/ which is long.

There are two kinds of phonetic-to-phoneme rules for vowels which must be included in the system. The first concerns vowel length. Some of the vowel phonemes can be realised as short vowels in certain contexts. Specifically:

			Example of context
[I]	→	/i/	preceding: /ng, k/
[U]	→	/u/	preceding: /ng, k/
[E]	→	/e/	preceding: /j/
[OE]	→	/oe/	preceding: /j, n, t/
[O]	→	/o/	preceding: /w/

The second major rule concerns phonetic diphthongs. One of the arguments for supposing that there are no phonemic diphthongs in Cantonese can be advanced on phonotactic grounds. Essentially, phonetic diphthongs such as [iu] never occur before tautosyllabic consonants. This can be explained by assuming that (i) Cantonese syllables are closed by at most one consonant and (ii) phonetic diphthongs are underlyingly a sequence of a vowel plus glide (thus e.g. [iun] is excluded because it would consist of two final consonants thereby violating the phonotactic constraints in the coda). Consequently, rules must be included to map the phonetic diphthongs onto a sequence of vowel plus glide, e.g:

[oey] → /oe j/
[au] → /a w/

FUTURE DEVELOPMENTS

mu+ has now been adapted both to French (and specifically to read multichannel articulatory data from the ACCOR speech database project; see Hardcastle *et al.*, 1992) and to Cantonese. Adapting mu+ to languages other than English has highlighted some of the ways in which the system should be made more flexible. Specifically, two major changes to the design of the system are in progress. These are summarised below.

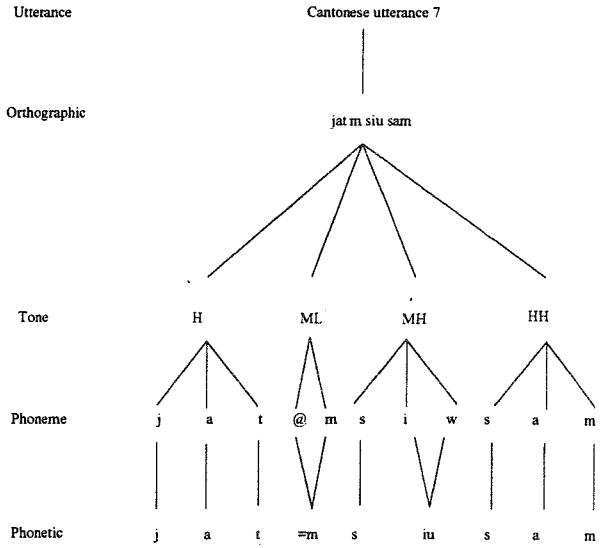

Fig. 1: Tree-structure for a single word utterance in Cantonese

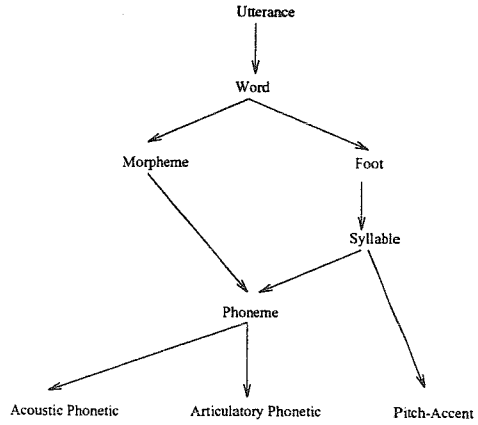

Fig. 2: One example of a possible structure relating different levels

Database Structure

Currently, mu+ supports databases with a single hierarchical structure (as in Fig. 1); however, some database applications require a more general structure than this. The new system will allow the database structure to be a Directed Acyclic Graph (DAG) restricted only by the requirement that it has a single top level node, usually called the utterance level. This allows us to encode 'overlapping' tree structures in a single database.

Fig. 2 shows a possible structure which combines a word based labelling which dominates foot, syllable and phoneme levels, with a morphemic level which dominates the same phoneme level. The phoneme level dominate two independent sets of lower level labels which are phonetic annotations of acoustic and articulatory data. Additionally, the syllable level dominates the pitch-accent level at which tone targets are marked.

This structure can be viewed via mu+ queries as a set of independent hierarchical descriptions. This allows what would normally be conflicting descriptions of an utterance to exist in the same database and share whatever label sets are common to each. It also will allow multiple physical labellings (Acoustic and articulatory Phonetic) to be integrated into one database.

Customisation

The new system is designed to be modular and to support the addition of new database creation modules as required. Two rule systems will be implemented to enable customisation of phonetic to phoneme conversion for different languages and to enable users to specify rules for creating new types of levels in the tree.

Each level of labels in the database will be created either automatically by a labelling module or by hand via the label editor. The editor will support the graphic manipulation of tree-like descriptions of the utterance and allow the addition and deletion of labels in the tree. Labelling modules will be implemented either in one of the provided rule languages (such as phonetic to phoneme rules or the more general label rewrite rules) or using the mu+ core library via an extension language (for example C++, perl or tcl). This modular approach should allow virtually any database structure to be generated automatically where this is possible.

Another approach to automatic labelling is a flexible dictionary lookup module. Dictionaries may contain entries for any label type (eg. words) and associate them with a set of lower-level labels (eg. syllables). The dictionary module constructs a new level of labelling given one level and the dictionary. This approach allows the use of existing dictionaries and provides an easy way of automating the construction of databases.

REFERENCES

Cheung K-H. (1986) The phonology of present day Cantonese. Unpublished Ph.D diss. University College: London.

Hardcastle W.J., Marchal A., Nicolaidis K., and Nguyen-Trong N. (1992) Non-linear annotation of multi-channel speech data. e. In Pittam J. (ed.) *Proc. of the 4th International Conference on Speech Science and Technology*, 718-722. University of Queensland Printery: Brisbane.

Harrington J., Cassidy S., Fletcher J. and McVeigh A. (1993) The mu+ system for corpus based speech research. *Computer Speech and Language*, 7, 305-331.

So Lydia K.H. & Thelwall Robin (1992) Hong Kong spoken Cantonese database. In Pittam J. (ed.) *Proc. of the 4th International Conference on Speech Science and Technology*, 542-547. University of Queensland Printery: Brisbane.

ACKNOWLEDGEMENTS

This research was supported by the Hong Kong Research Council grant HKU 242/92H and the Arthur Samy memorial fund.