Philology

Typesetting Chinese *pinyin* Using Virtual Fonts

Wai Wong

Pinyin is a phonetic system for transcribing the pronunciation (in Mandarin) of Chinese characters. It is in widespread use in China [1]. *Pinyin* is taught to millions of children in schools there. It is also used to teach foreigners and speakers of other Chinese dialects to learn Mandarin. This paper describes a package for typesetting *pinyin* in LaTeX [2]. It uses the features of virtual fonts and ligatures to provide an easy way of generating fonts for *pinyin* and a simple input method. This package requires a `.dvi` driver which is capable of handling virtual fonts.

1 What is *pinyin*

Pinyin uses the Latin alphabet to transcribe the sound of Chinese characters. All Chinese characters are single syllables. Each syllable can be divided into two parts: the *initial* and the *final*. The initials are consonants (or shēng mǔ in Chinese). Some of the finals (or yùn mǔ in Chinese) are pure vowels, the others are combinations of vowels and consonants. Each syllable can be pronounced in one of four possible tones (shēng diào). When writing down *pinyin*, the tone of the characters is indicated by placing a *tone mark* on top of the main vowel in the finals. For example the vowel a can have the following four tones ā á ǎ à.

There are two styles of writing *pinyin*: the first is to separate syllables by spaces, i.e., the *pinyin* for each character is separated, the second puts the *pinyin* of multi-character words together, like pīnyīn. When writing in the second form, syllables beginning with a, e or o may follow a vowel-ending syllable. This will cause confusion. If such case arises, an apostrophe mark (') is used to separate them. For example pí'aǒ (meaning leather jacket).

2 How to typeset *pinyin*

Since *pinyin* uses the Latin alphabet, it would not be too difficult to typeset it in LaTeX, except one has to place the tone marks properly. The normal LaTeX accent mark macros may be used for this purpose, but they are not very convenient since you need to put a tone mark on each syllable. (Try typing in a paragraph of properly tone-marked *pinyin* using those macros yourself.)

ā	a-	á	a'	ǎ	av	à	a‘
ē	e-	é	e'	ě	ev	è	e‘
ī	i-	í	i'	ǐ	iv	ì	i‘
ō	o-	ó	o'	ǒ	ov	ò	o‘
ū	u-	ú	u'	ǔ	uv	ù	u‘
ǖ	u:-	ǘ	u:'	ǚ	u:v	ǜ	u:‘
ü	u:	ê	e^				
Ā	A-	Á	A'	Ǎ	Av	À	A‘
Ē	E-	É	E'	Ě	Ev	È	E‘
Ī	I-	Í	I'	Ǐ	Iv	Ì	I‘
Ō	O-	Ó	O'	Ǒ	Ov	Ò	O‘
Ū	U-	Ú	U'	Ǔ	Uv	Ù	U‘
Ǖ	U:-	Ǘ	U:'	Ǚ	U:v	Ǜ	U:‘
Ü	U:	Ê	E^				

Table 1: Input formats of *pinyin* characters

The *pinyin* package provides a new style option `pinyin` to simplify the input and typesetting of *pinyin*. This option provides an environment, a simple input method and several fonts specially built for *pinyin*. To use it, put the word `pinyin` in the optional argument of the `\documentstyle` command. Then, within your document, the `pinyin` environment can be used anywhere.

Within the `pinyin` environment, tone marks are input by typing the appropriate character after the main vowel. The characters to mark the tones are: `- ' v ‘`. For example, `ma- ma' mav ma‘` will produce mā má mǎ mà. Another example shown below is a poem by the famous poet lǐ bái in the Tang dynasty. The input

```
\begin{quote}%
\begin{pinyin}%
chua'ng qia'n mi'ng yue‘ gua-ng\\
yi' shi‘ di‘ sha‘ng shua-ng\\
juv to'u wa‘ng mi'ng yue‘\\
di- to'u si- gu- xia-ng
\end{pinyin}%
\end{quote}
```

produces

chuáng qián míng yuè guāng
yí shì dì shàng shuāng
jǔ tóu wàng míng yuè
dī tóu sī gū xiāng

The input formats of all possible combinations of the tone marks and vowels are listed in Table 1.

	′0	′1	′2	′3	′4	′5	′6	′7	
′20x		Ā	Á	Ǎ	À	Ē	É	Ě	″8x
′21x	È	Ī	Í	Ǐ	Ì	Ō	Ó	Ǒ	″8x
′22x	Ò	Ū	Ú	Ǔ	Ù	Ǖ	Ǘ	Ǚ	″9x
′23x	Ǜ	Ü	Ê						″9x
′24x		ā	á	ǎ	à	ē	é	ě	″Ax
′25x	è	ī	í	ǐ	ì	ō	ó	ǒ	″Ax
′26x	ò	ū	ú	ǔ	ù	ǖ	ǘ	ǚ	″Bx
′27x	ǜ	ü	ê						″Bx
	″8	″9	″A	″B	″C	″D	″E	″F	

Table 2: *pinyin* character positions

The `pinyin` environment can take an optional argument which specifies the font to use. The currently available fonts are `sf` and `rm`. Conventionally, *pinyin* is printed in a Sans Serif font, so `sf` is the default as you can see in the examples above. If you say

```
\begin{pinyin}[rm] ... \end{pinyin}
```

the *pinyin* text will be typeset in a Roman font.

Very often, a small piece of *pinyin* is embedded in running text. A shorthand command `\py` can be used instead of the `pinyin` environment. For example, `\py{pi-n yi-n}` produces pīn yīn. Like the environment, `\py` can also take an optional argument to specify the font.

There is a command to specify the size of the *pinyin* characters: `\pysize`. It takes a single argument which is the required size. It can be any reasonable size for text, such as `10pt`, `11pt` and so on. The default is `10pt`.

3 The implementation of the `pinyin` package

The `pinyin` environment is implemented using virtual fonts and ligatures. Two families of virtual fonts, namely `pyr`*n* and `pyss`*n*, were created for typesetting *pinyin*. These *pinyin* fonts are needed because:

- accented characters, e.g., ǎ, ǖ, which are not included in ordinary TEX fonts and not even in the extended TEX (DC/EC) fonts, are required in *pinyin*;
- a special ligature table is required to support the input method in the `pinyin` environment.

These *pinyin* fonts are based on the Computer Modern fonts, i.e., `pyr10` is modified from `cmr10` and `pyss10` from `cmss10`. Ligature tables were defined for each of the vowels to map the combinations of the vowel and tone mark character to new character positions. All new characters and their codes are listed in Table 2[1]. The compound characters are defined in terms of the base vowel characters and the appropriate accent characters. These mappings and character definitions were done in the property list files. For example, the `pyss10` font is created by the following procedures:

1. convert `cmss10.tfm` to `cmss10.pl` using `tftopl`;
2. copy `cmss10.pl` into a file named `pyss10.vpl`;
3. edit `pyss10.vpl` to add the ligature tables and new character definitions;
4. generate `pyss10.tfm` and `pyss10.vf` using `vptovf`.

The ligature tables for all *pinyin* fonts should be identical. Appendix A lists the complete ligature table. When editing the `.vpl` file, care must be taken to merge this into the existing table in the file.

The new character definitions have a general form that looks like the example below:

```
(CHARACTER O 241
   (COMMENT a1 a-)
   (MAP (PUSH) (SETCHAR C a) (POP)
      (SETCHAR O 26))
   (CHARWD R 0.480557)
   (CHARHT R 0.608887)
   )
```

[1] The positions of the lowercase `pinyin` characters are arranged in a way that their codes are equal to the least significant byte of GB codes (the Chinese national standard). For example, ā at position `"A1` has GB code `A8A1` in hexadecimal.

for single *accent* in (¯, ´, ˇ, `, ¨, ˆ)
  if the base character *basechar* is in (a, e, o, u)
    wd is the width of *basechar*
    ht is the height of *accent*
    no movement is need for setting the character
    the program is:
    (PUSH) (SETCHAR *basechar*) (POP)
    (SETCHAR *accent*)
  if the base character *basechar* is i .
    basechar becomes ı (the dotless i)
    wd is the width of *basechar*
    ht is the height of *accent*
    the program is:
    (PUSH) (MOVELEFT $(wd_{accent} - wd_{basechar})/2$)
    (SETCHAR *accent*) (POP)
    (SETCHAR dotless i)
for double *accent*
    wd is the width of the base character (u)
    $ht \leftarrow ht_{ddot} + ht_{accent} - ht_{\mathrm{u}}$
    the program is:
    (PUSH) (SETCHAR C u) (POP)
    (PUSH) (SETCHAR O 177) (POP)
    (MOVEUP $ht_{ddot} - ht_{\mathrm{u}}$)
    (SETCHAR *accent*)

Figure 1: Algorithm for *pinyin* character definitions.

This is the definition for character ā whose character code is *'241*. The MAP property specifies how this character is generated. It first saves the current position, typesets the character a in the current base font and recovers the current position. It then overprints the accent character ¯ whose code is *'26*. This is the simplest case since the accents in CM fonts were designed so that they do not need to be raised or lowered for most lower case letters. For more complicated characters like ǖ, the second accent mark (the one on top) has to be moved up. The amount can be calculated using the heights of the base character and the accents given in the original `.pl` file. The CHARWD and CHARHT properties specify the width and height of the new character. They are calculated from the widths and heights of the base and accent characters.

The algorithm in Figure 1 was used in building the *pinyin* fonts where *wd* and *ht* are respectively the width and height of the new characters. This gives a precise specification of the new characters. The resulting appearance of the compound characters is very good though not perfect. The advantage of using an algorithm is that it could be possible to mechanize the generation of new font files by a program. To achieve the perfect result, manual editing, i.e., moving the accents or base character around, is necessary.

4 A sample installation

This package has been developed and tested in UNIX systems. It will not be difficult to port it to other systems; in fact, the style file and the font files can be moved to other systems without any change provided that the new environment has TEX and `.dvi` driver supporting virtual fonts. The remainder of this section describes an installation procedure in UNIX systems as an example.

This package is distributed as a `uuencoded compressed tar` archive file for UNIX users. Use the commands below to decode and extract the files:

```
uudecode pinyin.tar.Z.uue
zcat pinyin.tar.Z | tar xvf -
```

Afterward there will be a directory containing the style file `pinyin.sty` and the `.vf` and `.tfm` files. To install it, follow the procedures below:

1. copy `pinyin.sty` to the directory where TEX looks for macro files. In UNIX systems it is usually `/usr/local/lib/tex/inputs`.
2. copy all the `.tfm` files to the directory where TEX looks for fonts. In UNIX systems it is usually `/usr/local/lib/tex/fonts/tfm`.
3. copy all the `.vf` files to the directory where your `.dvi` driver looks for virtual fonts. For `dvips` in UNIX systems, it is usually `/usr/local/lib/tex/fonts/vf`.

5 Conclusions and future work

The `pinyin` package provides a simple and flexible means of typesetting *pinyin* text. Its usefulness is obvious to anyone who wants to perform this task. It is very easy to use. The output quality is high, thanks to the TEX system.

Although the *pinyin* package currently has a small number of fonts and they are hard-wired in the `pinyin` environment currently, it is sufficient for most applications since *pinyin* is only a phonetic transcription, it is usually mixed with ordinary text in a document and seldom appears as an entire *pinyin* document on its own. Actually, having special fonts will provide a visual distinction between ordinary text and *pinyin*.

A natural extension to this work will be to automate the generation of new fonts by implementing the algorithm in Figure 1 and incorporating the ligature table. This will provide us an easy means of creating new font when the need arises.

The way of implementing this package, namely by using virtual fonts and ligatures, can be generalized for other languages, such as Vietnamese, which have many accented and double accented characters. For such applications, the automatic font generation program is necessary, and style files which support more flexible font change and the new font selection scheme (NFSS) are also required.

References

[1] Scheme for the chinese phonetic alphabet. Issued by the State Council of the People's Republic of China.

[2] Leslie Lamport. *LATEX: A Document Preparation System.* Addison-Wesley, 1985.

[3] Tomas Rokicki. *DVIPS: a TEX driver.*

◇ Wai Wong
Computer Laboratory
University of Cambridge
New Museum Site
Pembroke Street
Cambridge CB2 3QG
U.K.
ww@cl.cam.ac.uk

A Ligature table for *pinyin* characters

This appendix lists the ligature table for the *pinyin* characters. It should be merged into the existing table in the base font.

```
(LABEL C A)
(LIG 0 55 0 201)
(LIG 0 47 0 202)
(LIG C v 0 203)
(LIG 0 140 0 204)
(STOP)
(LABEL C E)
(LIG 0 55 0 205)
(LIG 0 47 0 206)
(LIG C v 0 207)
(LIG 0 140 0 210)
(LIG 0 136 0 232)
(STOP)
(LABEL C I)
(LIG 0 55 0 211)
(LIG 0 47 0 212)
(LIG C v 0 213)
(LIG 0 140 0 214)
(STOP)
(LABEL C O)
(LIG 0 55 0 215)
(LIG 0 47 0 216)
(LIG C v 0 217)
(LIG 0 140 0 220)
(STOP)
(LABEL C U)
(LIG 0 55 0 221)
(LIG 0 47 0 222)
(LIG C v 0 223)
(LIG 0 140 0 224)
(LIG 0 072 0 231)
(STOP)
(LABEL 0 231)
(LIG 0 55 0 225)
(LIG 0 47 0 226)
(LIG C v 0 227)
(LIG 0 140 0 230)
(STOP)

(LABEL C a)
(LIG 0 55 0 241)
(LIG 0 47 0 242)
(LIG C v 0 243)
(LIG 0 140 0 244)
(STOP)
(LABEL C e)
(LIG 0 55 0 245)
(LIG 0 47 0 246)
(LIG C v 0 247)
(LIG 0 140 0 250)
(STOP)
(LABEL C i)
(LIG 0 55 0 251)
(LIG 0 47 0 252)
(LIG C v 0 253)
(LIG 0 140 0 254)
(STOP)
(LABEL C o)
(LIG 0 55 0 255)
(LIG 0 47 0 256)
(LIG C v 0 257)
(LIG 0 140 0 260)
(STOP)
(LABEL C u)
(LIG 0 55 0 261)
(LIG 0 47 0 262)
(LIG C v 0 263)
(LIG 0 140 0 264)
(LIG 0 072 0 271)
(STOP)
(LABEL 0 271)
(LIG 0 55 0 265)
(LIG 0 47 0 266)
(LIG C v 0 267)
(LIG 0 140 0 270)
(STOP)
```