CSME 206A Natural language & Speech Processing **Spring Semester**

Lecture 6: Computational Phonology

Lecturer: K.R. Chowdhary *: Professor of CS*

6.1 Mapping Text to phones for TTS

An important application of this is *pronunciation dictionaries*. These dictionaries are used for TTS (text-to-speech) systems. The simplest pronunciation dictionaries are to have a list of words and their pronunciations, like shown in table 6.1.

However, the challenge is due to large number of words, over one hundred thousand at the minimum. The other challenge is that two distinct words are many a times pronounced the same, called *homographs*. For example, the verb *wind* (You should wind up soon!) is pronounced [waind], while the noun *wind*, in "Tough winter wind", is pronounced as [wind]. Hence, the correct context should be supplied to TTS, to identify as which form of that word is right at that moment.

The other challenge is proper nouns (the name of person, or place, or company, etc). These account for millions in number.

Table 6.1: Mapping Text to phones.

Word	Pronunciation
cat	[kaet]
cats	[caets]
fox	[fax]
foxes	[fak.siz]
pigs	[pigz]
geese	[gis]

6.2 TTS Beyond the Lookup Dictionaries

The method discussed above is fine if all the words are built up in the dictionary, in advance. However, it is not possible to represent every word in dictionary, e.g., names of places, multilingual words, names of persons or things in different dialects, and numbers. Both the speech synthesis (TTS) and automatic speech recognition (ASR) systems need to be able to guess the pronunciation of words that are not in their dictionary.

6.2.1 FST-based Pronunciation Lexicons

Early work in pronunciation modeling to TTS relied heavily on letter to sound rules, where each rule specified how a letter or combination of letters was mapped to phones; as given in the table 6.2.

These systems comprised a long list of such rules and a very small dictionary of exceptions (often function words such as a, are, as , both, does, etc.).

Table 6.2: Text fragment v/s Pronunciation mapping.

Text Fragment	Pronunciation (Phone)
-p-	[p]
-ph-	[f]
-phe-	[fi]
-phes-	[fiz]
-place-	[pleis]
-placi-	[pleisi]
-plement-	[pliment]

More recent systems rely on very large dictionaries, with letter to sound rules only used for small number of words that are neither in the dictionary nor are morphological variants of words in dictionary.

The new approach, which is based on FST has following components:

1. An FST to represent pronunciation of individual words and morphemes in a lexicon (table 6.3). Note that it requires two tapes.
2. FST to represent possible sequence of morphemes,
3. Individual FST for each pronunciation rule (e.g., for -s in plural),
4. Letter to sound rules for names and acronyms. Note that it requires two tapes.
5. Default letter to sound rules for other unknown words.

The first of above is simple extension of what was discussed earlier. The extended version is given in the table 6.3, which corresponds to lexical tapes (top most) in figure 6.2. The arcs for regular and irregular nouns in figure 6.1 corresponds to the mapping shown in table 6.3, where, each symbol in lexicon is a pair of symbols separated by "|", the first representing "orthographic" lexical entry and second "phonological" lexical entry.

Table 6.3: Phonetic - Transcription

Orthographic-lexicon	Lexicon
Regular noun	
cat	$c\|k\ a\|ae\ t\|t$
fox	$f\|f\ o\|a\ x\|ks$
dog	$d\|d\ o\|a\ g\|g$
Irregular noun	
goose	$g\|g\ oo\|a\ s\|s\ e\|\varepsilon$

The FST shown in figure 6.1 is for nomial singular and plural inflection. The automata adds the morphological features [+N], [+PL], and [+SG] at the lexical level and also adds the plural suffix $s|z$ at the intermediate level.

The final FST is produced by combining the composition shown in table 6.3 along with the FST shown in figure 6.1. The figure 6.2 shows the mapping between the lexicon and surface form for orthography and

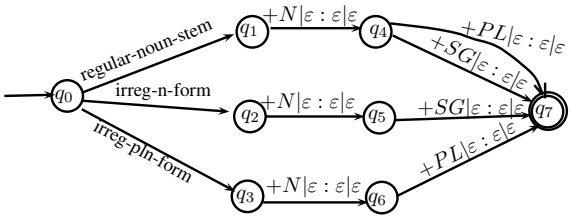

Figure 6.1: FST for text to speech.

phonology simultaneously, hence there are two tapes at every level, one for sound and other for spelling. The system can be used to map from a lexical entry to its surface pronunciation or from surface orthography to surface pronunciation via the lexical entry.

The FST figure 6.1 shows two level, an underlying or "lexical" level and Intermediate level. The only thing which remains to be added is transducers which apply the spelling rules; and pronunciation rules to map intermediate level into surface level. These include the spelling and pronunciation rules discussed in ε-insertion and e.g., flapping sound rules.

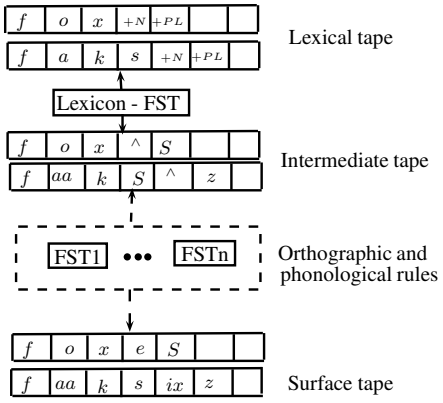

Figure 6.2: Mapping between lexicon and surface.

Recall that lexicon FST maps between the "lexical" level, with its stems and morphological features, and an "intermediate" level which represents a simple concatenation of morphemes. Then a host of FSTs, each representing either representing a single spelling rule constrainst, or a single morphological constrainst, all run in parallel so as to map between this intermediate level and the surface level.

Note that, each level has both orthographic and phonological representations. For text to speech applications, in which the input is lexical form (e.g., for text generation, where the system knows the identity of the word, its part-of-speech, its inflection, etc), the cascade of FTSs can map from lexical form to surface pronunciations. For text-to-speech applications in which the input is a surface spelling (e.g., for "reading text out loud" applications), the cascade of FTSs can map from surface orthographic form to surface pronunciation va underlying lexical form.

References

[DJJM02] D. Jurafsky and J. Martin, "Speech and Language Processing," *Pearson India*, 2002, Chapter 4.