Morphology and Rhyming: Two Powerful Alternatives to Letter-to-Sound Rules for Speech Synthesis

Cecil H. Coker
Kenneth W. Church
Mark Y. Liberman

AT&T Bell Laboratories
600 Mountain Ave.
Murray Hill, N.J., 07974

Abstract

Most speech synthesizers have tended to depend on letter-to-sound rules for most words, and resort to a small ‘‘exceptions dictionary’’ of about 5000 words to cover the more serious gaps in the letter-to-sound rules. The Bell Laboratories Text-to-Speech system, *TTS*, takes a radical dictionary-based approach; dictionary methods (with morphological and analogical extensions) are used for the vast majority of words. Only a fraction of a percent (0.5% of words overall; 0.1% of lowercase words) are left for letter-to-sound rules. Moving to an extreme dictionary-based approach cuts the error rate by at least an order of magnitude. Now that the dictionary is the rule and not the exception, the term ‘‘exceptions dictionary’’ seems somewhat dated.

1. Background

A speech synthesizer is a machine that inputs a stream of text and outputs a speech signal. This paper will discuss a small piece of this problem, the conversion of words into phonemes. Typically, the conversion is accomplished in one of two ways: either (1) by looking the words up in a dictionary (with possibly some limited morphological analysis), or (2) by sounding the words out from their spelling using basic principles.

Both approaches have their advantages and disadvantages. Most speech synthesizers adopt a hybrid strategy: employing letter-to-sound rules for most words, and catching the most common irregular words with a small ‘‘exceptions dictionary’’ of 5,000 words or less. MITalk took a radical dictionary-based approach for its day. A dictionary of 10,000 morphemes (Allen, Hunnicutt, Klatt, 1987, p. 25) covered the vast majority of the input words. Only 5% of the input words could not be handled by the *decomp* module and had to be passed to Hunnicutt’s letter-to-sound rules (Allen, personal communication).

2. The dictionary is the rule, not the exception

The Bell Laboratories Text-to-Speech system, *TTS*, takes an even more radical dictionary-based approach; dictionary methods are used for 99.9% of the input words, and only the remaining 0.1% will be passed to *namsa*, a letter-to-sound rule system designed for surnames (Church, 1986).

The main motivation for moving to an extreme dictionary-based approach is accuracy. In general, table lookup is more accurate than starting from basic principles (e.g., letter-to-sound rules). Dictionary-based systems make only a few errors in 10,000 ordinary words (names and typos are harder). In contrast, a good letter-to-sound system such as Hunnicutt’s (Allen, Hunnicutt, Klatt, 1987, chapter 6) will badly misprounce 10-20% of these words. Self-organizing/connectionist systems such as (Sejnowski and Rosenberg, 1987) are so much worse that they report performance by letter rather than by word. (Their error rate of 10-20% by letter corresponds to a word error rate of approximately 50%.) In short, dictionary methods are much more accurate than letter-to-sound rules.

In the early days of speech synthesis, the dictionary-based approach faced two problems: memory and coverage. The memory problem has been much alleviated with declining memory prices, though for some applications (e.g., hand-held talking dictionaries), 1/4 megabytes is still a minor concern. Coverage is a more serious issue, especially for surnames, which are generally thought to be more difficult than ordinary words that one might find in a collegiate dictionary. There are many more names; a list of surnames supplied by Donnelley Marketing contains 1.5 million types (72 millions tokens), considerably larger than the number of words in an unabridged dictionary (0.25 million types). In addition, it takes a much larger list of names to achieve a given level of coverage. For example, to cover half of the surnames of in the United States requires a list of more than 2300 names, whereas, for ordinary words, the same 50% level of coverage can be achieved with a dictionary of only 141 words. And, names are thought to be less amenable to derivation techniques. Names come from many different languages; the methods of derivation are more diverse and language-specific.

David Schulz and Beth Schulz (AT&T Bell Laboratories, Indian Hill Park) have recently constructed a dictionary of the 50,000 most frequent surnames in the United States so that it is no longer necessary to use the letter-to-sound system *namsa* for these names. This greatly improves performance on a corpus of names such as the Kansas City Telephone Book. It is believed that *namsa* by itself produces good results 50% of the time and acceptable results about 85% of the time (David Schulz, personal communications). The dictionary itself covers 87% of Kansas City. Thus, if *namsa* is somewhere between 50% and 85% correct by frequency, then the combination of the dictionary plus *namsa* should yield between 87% + (50%)(13%) = 93.5% and 87% + (85%)(13%) = 98% performance, a significant improvement over *namsa* alone.

3. Morphological and Analogical Extensions

The argument becomes considerably stronger when we consider morphology and analogical extensions to the dictionary. Names such as *Walters* and *Lucasville* can be derived from other names by very simple morphological processes. These stress-neutral processes increase the coverage of the names dictionary by 25%, as indicated in Table 1.

More complicated stress shifting processes such as *Jordan* → *Jordanian* and *Washington* → *Washingtonian* have also been implemented. One might note, with some disappointment, that these more complicated processes do not produce great benefits in coverage. Table 2 shows that stress shifting morphology (e.g., primary-stress endings, suffix-exchange,[1] *ity*-class endings, *al*-class endings) contributes considerably less than stress neutral morphology.

Sometimes surprisingly simple processes provide the greatest benefits. The rhyme analogy method is one such case. The pronunciation of an unknown name such as *Plotsky* is determined by analogy with *Trotsky*, which happens to be in the names dictionary. The pronunciation of *Plotsky* is computed from the pronunciation of *Trotsky* by removing the initial /tr/ of *Trotsky* and replacing it with /pl/. It is remarkable just how many names can be pronounced in this way. As Table 1 shows, the rhyme method covers more names than many of the more complicated morphological processes.

There is, of course, some chance of error with the rhyme analogy method. For example, we wouldn't want to derive *Jose* from *hose*. It is not possible to know for sure if two words rhyme by looking at their spelling alone. The heuristic employed by the rhyme analogy method is correct about 90% of the time. Although far from perfect, this heuristic is more reliable than letter-to-sound rules. Given a choice between the rhyming heuristic and letter-to-sound rules, it is much safer to choose the rhyming heuristic.

The rhyming method uses letter-to-sound rules in a relatively safe region (the beginning of the word) and uses dictionary methods for the rest of the word. Another method, which we call *Interior Letter-to-Sound*, also combines letter-to-sound rules and dictionary methods. This method uses letter-to-sound rules for an interior syllable such as the *-ar-* in *Daddario*. The remainder of *Daddario* is dervied from *Addonizio*, which happens to be in the dictionary, by suffix-exchange and rhyming. Suffix exchange is used to replace the *-izio* in *Addonizio* with *-io*; rhyming is used to add the initial *d-*. Thus we use Interior Letter-to-Sound as an extension of ending exchange, to infer pronunciation for a syllable preceeding a recognized stress-forcing ending. The ending tells whether the preceeding syllable will be unstressed or primary stressed; and if

1. We introduce the term *suffix-exchange* to refer to a process (like Aronoff's truncation operation) of substituting one affix for another (in the same class) such as *nominate* → *nominee.*

stressed, whether the vowel will be tense or lax.

4. Coverage

Table 1 gives the coverage for the 1/4 million most frequent names in the Donnelly Marketing List. Although the dictionary-based methods cover a large percentage of names, they do not always produce the right pronunciation. Even the direct-hit method makes a few errors since the surnames dictionary was constructed quickly under considerable time pressure. (In the future, we hope to have the surname dictionary corrected by a team of professional lexicographers.)

The results of an informal evaluation by a single human judge, Jill Burstein, are presented in Table 1. The judge listened to almost 1000 names and graded them on a 3-way scale: (1), good (''that's the way I would have said it''), (2), OK/? (''I probably wouldn't say it that way, but I could imagine someone else doing so'' or ''I'm not sure know how it should be said''), and (3), poor (''I know that's wrong''). This evaluation shows that compounding is considerably more risky than the other processes (because compounding combines two stems whereas most other processes deal with just a single stem). In general, surnames are very hard; the error rate for ordinary words are much smaller.

Table 2 shows the coverage of the various methods for words distributed over the Associated Press Newswire during 1988. This corpus is very different than the Donnelly list of surnames. There are a large number of uppercase words in the AP corpus, only some of which are names. For the purposes of this paper, a word is considered to be a name if it appears in uppercase at least one hundred times more often than it appears in lowercase.

5. Conclusion

The pronunciation problem has traditionally been divided into two very separate modules: letter-to-sound rules and the exceptions dictionary. The focus has been on letter-to-sound rules which work from first principles. In contrast, the present work resorts to letter-to-sound rules only when all alternatives have been exhausted. The most reliable inference is table lookup. Failing that, the system tries to make as safe an inference as possible from similar words in the dictionary. Stress neutral morphology is considered fairly safe; rhyming is more dangerous, but far more reliable than letter-to-sound rules. Our approach breaks down the traditional barriers between letter-to-sound rules and dictionary-based methods. The rhyme method, for example, uses letter-to-sound rules to pronounce the initial consonant onset and dictionary methods to pronounce the remainder. The interior letter-to-sound method is a more ambitious hybrid approach.

Examples of dictionary extensions

stress-neutral ending:

abandons = abandon + s
abandoning = abandon + ing
abandonment = abandon + ment
Abbotts = Abbott + s
Abelson = Abel + son

primary-stress ending:

addressee = address + ee
abductee = abduct + ee
accountability = account + ability
activization = active + ization
adaptation = adapt + ation

ity-**class ending:**

abortion = abort + ion
abnormality = abnormal + ity
academician = academic + ian
Adamovich = Adam + ovich
Ambrosian = Ambrose + ian

al-**class ending:**

accidental = accident + al
adjectival = adjective + al
combative = combat + ive

suffix-exchange:

nominee = nominate – ate + ee
auditoria = auditorium – um + a
collusive = collude – ude + usive
eldress = elder – er + ress
Agnano = Agnelli – elli + ano;
Bierstade = Bierbaum – baum + stadt

prefix:

adjoin = ad + join
cardiovascular = cardio + vascular
chlorofluorocarbon = chloro + flurocarbon
O'brien = O' + brien
Macdonald = Mac + donald

compound:

airfield = air + field
anchorwoman = anchor + woman
armrest = arm + rest
Abdulhussein = Abdul + hussein
Baumgaertner = Baum + gaertner

Rhyming:

Plotsky (from *Trotsky*)
Alifano (from *Califano*);
Anuszewski (from *Januszewski*)

This approach has a much smaller error rate than previous letter-to-sound systems.

References

Allen, J., Hunnicutt, M., and Klatt, D., ‘‘From text to speech: The MITalk system,’’ Cambridge University Press, 1987.

Aronoff, M., ‘‘Word Formation in Generative Grammar,’’ MIT Press, 1976.

Church, K., ‘‘Stress Assignment in Letter to Sound Rules for Speech Synthesis,’’ ICASSP 1986.

Sejnowski, T., and Rosenberg, C., ‘‘Parallel networks that learn to pronounce English text,’’ *Complex Systems*, vol. 1, pp. 144-168, 1987.

Table 1: Methods used for Names in Donnelly Marketing List

Method	Raw Counts		Percentage		Evaluation		
	Type	Token	Type	Token	Good	OK/?	Poor
Direct Hit	40,208	40,354,563	16.08%	59.34%	95	3	2
Name + Stress-Neutral Ending	42,454	16,996,558	16.98%	24.99%	98	6	4
Name + Primary-Stress Ending	627	69,016	.25%	.10%	94	6	6
Name + *ity*-class Ending	6,760	984,662	2.70%	1.45%	127	4	8
Name + *al*-class Ending	2,393	341,301	.96%	.50%	90	8	6
Name + Name (Compound)	16,619	1,663,444	6.65%	2.45%	76	3	13
Name + Suffix-Exhange	14,523	1,010,813	5.81%	1.49%	101	5	3
Rhyme with Name	29,924	1,579,499	11.97%	2.32%	84	8	4
Interior Letter-to-Sound	13,796	500,923	5.52%	.74%	71	8	5
Prefix	1,230	117,857	.49%	.17%			
Combinations of Above	43,874	3,023,149	17.55%	4.44%			
All Dictionary-based Methods	212,408	66,641,785	84.96%	97.99%			
Remainder (for *namsa*)	37,592	1,364,118	15.04%	2.01%			
Totals	250,000	68,005,903	100.00%	100.00%			

Table 2: Methods used for Words in 1988 Associated Press Newswire

Method	Ordinary Words (non-names) Raw Counts		Ordinary Words (non-names) Percentage		Capitalized Words (names) Raw Counts		Capitalized Words (names) Percentage	
	Type	Token	Type	Token	Type	Token	Type	Token
Direct Hit	13,356	20,646,656	18.27%	75.13%	26,965	4,147,321	22.87%	70.24%
Stress-Neutral	24,050	4556078	32.90%	16.58%	25,638	811,751	21.75%	13.75%
Primary-Stress	679	76,222	.93%	.28%	433	17,128	.37%	.29%
ity-Class	1,943	332,587	2.66%	1.21%	2,209	96,469	1.87%	1.63%
al-Class	1,174	237,979	1.61%	.87%	1,119	67,817	.95%	1.15%
Suffix-Exchange	497	36,763	.68%	.13%	2,409	23,356	2.04%	.40%
Rhyme					6,888	137,054	5.84%	2.32%
Prefix	2,907	436,839	3.98%	1.59%	780	12,945	.66%	.22%
Interior L-to-S					5,133	23,833	4.35%	.40%
Compound	3,718	170,877	5.09%	.62%	5,591	79,753	4.74%	1.35%
Combinations	15,151	966,033	23.46%	3.51%	32,705	496,026	27.76%	8.40%
All Methods	63,475	27,460,034	89.58%	99.92%	97,849	5,752,566	83.01%	97.43%
Remainder	9,618	21,076	10.42%	.06%	20,032	151,951	16.99%	2.57%
Totals	73,093	27,481,110	100.00%	100.00%	117,881	5,904,517	100.00%	100.00%