Extending Pronunciation Lexicons via Non-phonemic Respellings

Lucian Galescu
Florida Institute for Human and Machine Cognition
40 S Alcaniz St., Pensacola FL 32502, USA
lgalescu@ihmc.us

Abstract

This paper describes work in progress towards using non-phonemic respellings as an additional source of information besides spelling in the process of extending pronunciation lexicons for speech recognition and text-to-speech systems. Preliminary experimental data indicates that the approach is likely to be successful. The major benefit of the approach is that it makes extending pronunciation lexicons accessible to average users.

1 Introduction

Speech recognition (SR) systems use pronunciation lexicons to map words into the phoneme-like units used for acoustic modeling. Text-to-speech (TTS) systems also make use of pronunciation lexicons, both internally and as "exception dictionaries" meant to override the systems' internal grapheme-to-phoneme (G2P) convertors. There are many situations where users might want to augment the pronunciation lexicons of SR and TTS systems, ranging from minor fixes, such as adding a few new words or alternate pronunciations for existing words, to significant development efforts, such as adapting a speech system to a specialized domain, or developing speech systems for new languages by bootstrapping from small amounts of data (Kominek *et al*., 2008).

Unfortunately, extending the pronunciation lexicon (PL) is not an easy task. Getting expert help is usually impractical, yet users have little or no support if they want to tackle the job themselves. Where available, the user has to either know how to transcribe a word's pronunciation into the application's underlying phone set, or, in rare cases, use pronunciation-by-orthography, whereby word pronunciations are respelled using other words (e.g., "*Thailand*" is pronounced like "*tie land*"). The former method requires a certain skill that is clearly beyond the capabilities of the average user; the latter is extremely limited in scope.

What is needed is a method that would make it easy for the users to specify pronunciations themselves, without requiring them to be or become expert phoneticians. In this paper we will argue – with backing from some preliminary experiments – that non-phonemic respelllings might be an accessible intermediate representation that will allow speech systems to learn pronunciations directly from user input faster and more accurately.

2 Extending pronunciation lexicons

Automatic G2P conversion seems the ideal tool to help users with PL expansion. The user would be shown a ranked list of automatically derived pronunciations and would have to pick the correct one. To make such a system more user-friendly, a synthesized waveform could also be presented (Davel and Barnard, 2004; also Kominek *et al*., 2008). This approach has a major drawback: if the system's choices are all wrong – which is, in fact, to be expected, if the number of choices is small – the user would have to provide their own pronunciation by using the system's phonetic alphabet. In our opinion this precludes the approach from being used by non-specialists.

Other systems try to learn pronunciations only from user-provided audio samples, via speech recognition/alignment (Beaufays *et al*., 2003; see also Bansal *et al*., 2009 and Chung *et al*., 2004). In such systems G2P conversion may be used to constrain choices, thereby overcoming the notoriously poor phone-level recognition performance. For example, Beaufays *et al*. (2003) focused on a directory assistance SR task, with many out-of-vocabulary proper names. Their procedure works by initializing a hypothesis by G2P conversion, and thereafter refining it with hypotheses from the joint alignment of phone lattices obtained from audio samples and the current best hypothesis. Several transformation rules were employed to expand the search space of alternative pronunciations.

While audio-based pronunciation learning may appear to be more user-friendly, it actually suffers from being a slow approach, with many audio samples being needed to achieve reasonable performance (the studies cited used up to 15 samples). It is also unclear whether the pronunciations learned are in fact correct, since the approach was mostly used to help increase the performance of a SR system. The SR performance improvements (ranging from 40% to 74%) must be due to better

Proceedings of NAACL HLT 2009: Short Papers, pages 129–132,
Boulder, Colorado, June 2009.

pronunciations, but we are not aware of the existence of any correctness evaluations.

3 Non-phonemic respellings

The method proposed here is aimed at allowing users to directly indicate the pronunciation of a word via *non-phonemic respellings* (NPRs). With NPRs, a word's pronunciation is represented according to the ordinary spelling rules of English, without attempting to represent each sound with a unique symbol. For example, the pronunciation of the word *phoneme* could be indicated as \FO-neem\, where capitalization indicates stress (boldface, underlining, and the apostrophe are also used as stress markers). It is often possible to come up with different respellings, and, indeed, systematicity is not a goal here; rather, the goal is to convey information about pronunciation using familiar spelling-to-sound rules, with no special training or tables of unfamiliar symbols.

NPRs are used to indicate the pronunciation of unfamiliar or difficult words by news organizations (mostly for foreign names), the United States Pharmacopoeia (for drug names), as well as countless interest groups (astronomy, horticulture, philosophy, etc.). Lately, Merriam-Webster Online[1] has started using NPRs in their popular Word of the Day[2] feature. Here is a recent example:

girandole • \JEER-un-dohl\

While NPRs seem to be used by a fairly wide range of audiences, we mustn't assume that most people are familiar with them. What we do know, however, is that people can learn new pronunciations faster and with fewer errors from NPRs than from phonemic transcriptions and this holds true whether they are linguistically-trained or not (Fraser, 1997). We contend, based on preliminary observations, that not only are NPRs easily decoded, but people seem to be able to produce relatively accurate NPRs, too.

4 Our Approach

Our vision is that speech applications would employ user-provided NPRs as an additional source of information besides orthography, and use dedicated NPR-to-pronunciation (N2P) models to derive hypotheses about the correct pronunciation.

However, before embarking on this project, we ought to answer three questions:

1. Is generic knowledge about grapheme-to-phoneme mappings in English sufficient to decode pronunciation respellings? Or, in technical terms, are generic G2P models going to work as N2P models?
2. Are pronunciation respellings useful in obtaining the correct pronunciation of a word beyond the capabilities of a G2P converter?
3. Since we don't require that average users learn a respelling system, are novice users able to generate useful respellings?

In the following we try to answer experimentally the technical counterparts of the first two questions, and report results of a small study designed to answer the third one.

4.1 Data and models

We collected a corpus of 2730 words with a total of 2847 NPR transcriptions (some words have multiple NPRs) from National Cancer Institute's Dictionary of Cancer Terms.[3] The dictionary contains over 4000 medical terms. Here are a couple of entries (without the definitions):

lactoferrin (LAK-toh-fayr-in)
valproic acid (val-PROH-ik A-sid)

Of the 2730 words, 1183 appear in the CMU dictionary (Weide, 1998) – we'll call this the ID set. Of note, about 180 words were not truly in-dictionary; for example, *Versed* (a drug brand name), pronounced \V ER0 S EH1 D\, is different from the in-dictionary word *versed*, pronounced \V ER1 S T\. We manually aligned all NPRs in the ID set with the phonetic transcriptions.

We transcribed phonetically another 928 of the words – we'll call this the OOD set – not found in the CMU dictionary; we verified the phonetic transcriptions against the Merriam-Webster Online Medical Dictionary and the New Oxford American Dictionary (McKean, 2005).

For G2P conversion we used a joint 4-gram model (Galescu, 2001) trained on automatic alignments for all entries in the CMU dictionary. We note that joint n-gram models seem to be among the best G2P models available (Polyakova and Bonafonte, 2006; Bisani and Ney, 2008).

4.2 Adequacy of generic G2P models

To answer the first question above, we looked at whether the generic joint 4-gram G2P model is adequate for converting NPRs into phonemes.

At first, it appeared that the answer would be negative. We found out that NPRs use GP correspondences that do not exist or are extremely rare in the CMU dictionary. For example, the <[*ih*], \IH\> correspondence is very infrequent in the

[1] http://www.merriam-webster.com
[2] http://www.merriam-webster.com/cgi-bin/mwwod.pl
[3] http://www.cancer.gov

CMU dictionary (and appears only in proper names, e.g., *Stihl*), but is very frequently used in NPRs. Therefore, for the [*ih*] grapheme the G2P converter prefers \IH HH\ to the intended \IH\. Similar problems happen because of the way some diphones are transcribed. Two other peculiarities of the transcription accounted for other errors: a) always preferring /S/ in plurals where /Z/ would be required, and b) using [*ayr*] to transcribe \EH R\, which uses the very rare <[*ay*], \EH\> mapping. These deviations from ordinary GP correspondences occur with regularity and therefore we were able to fix them with four post-processing rules. We are confident that these rules capture specific choices made during the compilation of the Dictionary of Cancer Terms, to reduce ambiguity, and increase consistency, with the expectation that readers would learn to make the correct phonological choices when reading the respellings.

Another issue was that the set of GP mappings used in NPRs was extremely small (111) compared to the GP correspondence set obtained automatically from the CMU dictionary (1130, many of them occurring only in proper names). However, it turns out that 47524 entries in the CMU dictionary (about 45%) use exclusively GP mappings found in NPRs! This suggests that, while the generic G2P model may not be adequate for the N2P task, the GP mappings used in NPRs are sufficiently common that a more adequate N2P model could be built from generic dictionary entries by selecting only relevant entries for training. Unfortunately we don't have a full account of all "exotic" entries in the CMU dictionary, but we expect that by simply removing from the training data the approximately 54K known proper names will yield a reasonable starting point for building N2P models.

4.3 NPR-to-pronunciation conversion

To assess the contribution of NPR information to pronunciation prediction, we compare the performance of spelling-to-pronunciation conversion (the baseline) to that of NPR-to-pronunciation conversion, as well as to that of a combined spelling and NPR-based conversion, which is our end goal.

For the N2P task, we trained two joint 4-gram models: one based on the aligned NPRs, and a second based on the 47K CMU dictionary entries that use only GP mappings found in NPRs. Then, we interpolated the two models to obtain an NPR-specific model (the weights were not optimized for these experiments), which we'll call the N2P model. The combined, spelling and NPR-based model was an oracle combination of the G2P and the N2P model. Phone error rates (PER) and word error rates (WER) for both the ID set and the OOD set are shown in Figures 1 and 2, respectively. We

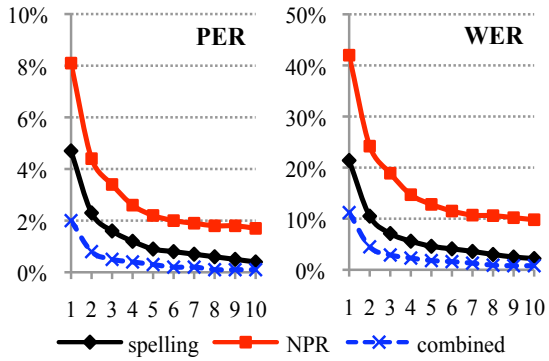

Figure 1. Phone and word error rates on the ID set.

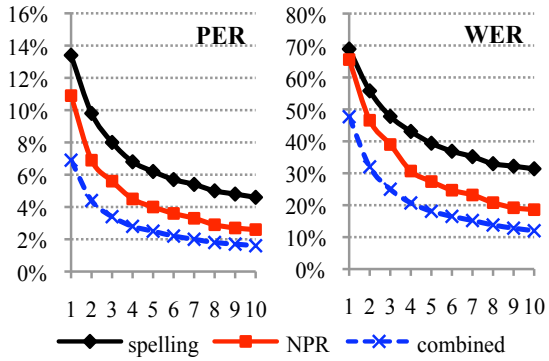

Figure 2. Phone and word error rates for the OOD set.

obtained n-best pronunciations with n from 1 to 10 for the three models considered.

As expected, G2P performance is very good on the ID set, since the test data was used in training the G2P model. Significantly, even though the N2P model is not as good itself, the combined model shows marked error rate reductions: for the top hypothesis it cuts the PER by over 57%, and the WER by over 47% when compared to the G2P performance on spelling alone.

Since the OOD set represents data unseen by either the spelling-based model or the NPR-based model, all models' performance is severely degraded compared to that on the ID set. But here we see that NPR-based pronunciations are already better than spelling-based ones. For the top hypothesis, compared to the performance of the G2P model alone, the N2P model shows almost 19% better PER, and almost 5% better WER, whereas the combined model achieves 49% better PER and close to 31% better WER.

4.4 User-generated NPRs

To answer the third question, we collected user-generated NPRs from five subjects. The subjects were all computer-savvy, with at least a BSc degree. Only one subject expressed some familiarity with NPRs (but didn't generate better NPRs than

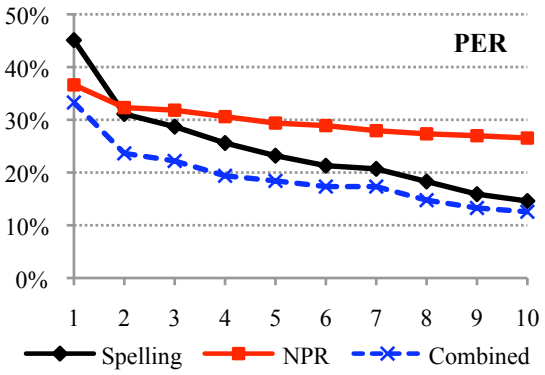

Figure 3. Phone error rates for user-generated NPRs.

other subjects).

The subjects were shown four examples of NPRs; two of them were recent Word of the Day entries, and had audio attached to them. The other two were selected from the OOD set. With only four words and two different sources we wanted to ensure that users would not be able to train themselves to a specific system. Subjects understood the problem easily and rarely if ever looked back at the examples during the actual test.

The test involved generating NPRs for 20 of the most difficult words for our generic GP model from the OOD set (e.g., *bronchoscope*, *parenchyma*, etc.). These words turned out to be mostly unfamiliar to users as well (the average familiarity score was just under 1.9 on a 4-point scale. No audio and no feedback were given.

Users varied greatly in the choices they made. For the word *acupressure*, the first two syllables were transcribed as AK-YOO in the Dictionary of Cancer Terms, and users came up with ACK-YOU, AK-U, and AK-YOU. This underscores that a good N2P model would have to account for far more GP mappings than the 111 found in our data.

Sometimes users had trouble assigning consonants to syllables (syllabification wasn't required, but subjects tried anyway), on occasion splitting them across syllable boundaries (e.g., \BIL-LIH-RUE-BEN\ for *bilirubin*), which guarantees an insertion error. It is quite likely that some error model might be required to deal with such issues.

Nonetheless, even though imperfect, the resulting NPRs showed excellent promise. Looking just at the top hypothesis, whereas the average PER on those 20 words was about 45% for the G2P model, pronunciations obtained from NPRs using the same G2P model (new GP mappings precluded the use of the N2P model described in the previous section) had only around 36% (+/-5%) phone error rate. The combined model showed an even better performance of about 33% (+/-5%) PER. Full results for n-best lists up to n=10 are shown in Figure 3.

5 Conclusions and Further Work

The experiments we conducted are preliminary, and most of the work remains to be done. More data need to be collected and analyzed before good NPR-to-pronunciation models can be trained. Further investigations need to be conducted to assess the average users' ability to generate NPRs and how they tend to deviate from the general grapheme-to-phoneme rules of English.

Nonetheless, we believe these experiments give strong indications that NPRs would be an excellent source of information to improve the quality of pronunciation hypotheses generated from spelling. Moreover, it appears that novice users don't have much difficulty generating useful NPRs on their own; we expect that their skill would increase with use. Particularly useful would be for the system to be able to provide feedback, including generating NPRs; we have started investigating this reverse problem, of obtaining NPRs from pronunciations, and are encouraged by the initial results.

References

D. Bansal, N. Nair, R. Singh, and B. Raj. 2009. A Joint Decoding Algorithm for Multiple-Example-Based Addition of Words to a Pronunciation Lexicon. *Proc. ICASSP'2009*, pp. 2104-2107.

F. Beaufays, et al. 2003. Learning Linguistically Valid Pronunciation From Acoustic Data. *Proc. Eurospeech'03*, Geneva, pp. 2593-2596.

M. Bisani and H. Ney. 2008. Joint-Sequence Models for Grapheme-to-Phoneme Conversion. *Speech Communication*, 50(5):434-451.

G. Chung, C. Wang, S. Seneff, E. Filisko, and M. Tang. 2004. Combining Linguistic Knowledge and Acoustic Information in Automatic Pronunciation Lexicon Generation. *Proc. Interspeech'04*, Jeju Island, Korea.

M. Davel and E. Barnard. 2004. The Efficient Generation of Pronunciation Dictionaries: Human Factors during Bootstrapping. *Proc. INTERSPEECH 2004*, Korea.

H. Fraser. 1997. Dictionary pronunciation guides for English. *International Journal of Lexicography*, 10(3), 181-208.

L. Galescu and J. Allen. 2001. Bi-directional Conversion Between Graphemes and Phonemes Using a Joint N-gram Model. *Proc. 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Scotland.

J. Kominek, S. Badaskar, T. Schultz, and A. Black. 2008. Improving Speech Systems Built from Very Little Data. *Proc. INTERSPEECH 2008*, Australia.

E. McKean (ed.). 2005. *The New Oxford American Dictionary* (2nd ed.). Oxford University Press.

T. Polyakova and A. Bonafonte, 2006. Learning from Errors in Grapheme-to-Phoneme Conversion. *Proc. ISCLP'2006*. Pittsburgh, USA.

R.L. Weide. 1998. The CMU pronunciation dictionary, release 0.6. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.