LEXICAL TUNING BASED ON TRIPHONE CONFIDENCE ESTIMATION

K.L. Markey
Berdy Medical Systems
4909 Pearl East Circle, Suite 202
Boulder, Colorado, USA 80301
Tel. 303-417-1603, FAX 303-417-1662, E-mail: markey@berdy.com

W. Ward
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA, USA 15213
Tel. 303-442-8807, FAX 303-417-1662, E-mail: whw@cs.cmu.edu

ABSTRACT

We propose and test a practical means of finding poor pronunciations and missing variants for large lexicons. We do so by statistically assessing the confidence of each phone in each pronunciation and comparing it with the statistical distribution of the same confidence metric for corresponding phones over the entire training corpus. A phone is targeted for correction for each word in which its mean score is significantly less than the phone's mean score over the entire training corpus. Neighboring phones are also reviewed for their contribution to the target phone's poor score. Thus far, we have experimented with this technique by manually correcting the pronunciation. In experiments with Wall Street Journal and dictated physical examination corpora, word error rates were reduced commensurate with the number of dictionary entries whose pronunciations were corrected as result of this process.

1. INTRODUCTION

Large pronunciation dictionaries are often created from diverse sources. Even when different labeling schemes have been mapped together, there are likely to be many pronunciation errors in the dictionary and many more suboptimal pronunciations. Even if technically correct from a linguistic perspective, some pronunciations could be suboptimal considering the actual inventory of trained phone models available to the system. Poor pronunciations generally account for a significant percentage of errors in newly developed systems. Correction is too labor intensive to be done by hand. Problems are compounded by the need to include dialectical and prosodic variants. These issues have spawned several responses, including the use of rules that model phonetic variability [1] or hidden Markov models that learn the patterns of phonetic variability [2,3]. These additional models add considerable complexity to the decoding process. They also fail to address the practical problem of how to find bad phonetic transcriptions in existing large dictionaries.

We propose and test a practical means of finding poor pronunciations and discovering missing variants for large lexicons. We do so by statistically assessing the confidence of each phone in each pronunciation and comparing it with the statistical distribution of the same confidence metric for corresponding phones over the entire training corpus. A phone is targeted for correction for each word in which its mean score is significantly less than the phone's mean score over the entire training corpus. Neighboring phones are also reviewed for their contribution to the target phone's poor score. At present we have experimented with this technique by manually correcting the pronunciation, but the correction is also being automated.

2. METHOD

To obtain acoustic scores, we use the Sphinx-II HMM-based speech decoder [4]. The decoder is used to produce a forced alignment of the acoustic training data to the corresponding reference transcripts. It chooses among alternative word pronunciations and determines the time alignment of each word, context-dependent phone, and state. The observation probability of each state is computed given the associated acoustic input. The acoustic score for each phone is derived by averaging probabilities over all corresponding states. For each observed triphone (context dependent phone), the mean and variance of the average acoustic score for the triphone is computed across all instances of the triphone in the training corpus. This is the population mean. In addition, we compute the mean acoustic score for each triphone in each lexical entry over all observed occurrences of that word in the data. This gives the absolute score for each phone in each word. To determine the score relative to the population mean, we compute the z-score relative to the population. The *z-score* for each phone *p*, in each word *w* is given by

$$z\text{-}score[w,p] = (\text{MEAN}[p] - score[w,p]) / \text{SD}[p]$$

where MEAN[*p*] is the mean and SD[*p*] is the standard error over all instances of phone *p*, and *score*[*w,p*] is the

mean over all instances of phone *p* in word *w*. For the results reported in this paper, we targeted all phones whose *z-score* was less than -2.0, that is all whose mean was more than 2 standard deviations less than the population mean.

3. EVALUATION

We ran the time-align Sphinx-II decoder on Wall Street Journal (SI-284) data for male speakers and determined which pronunciations were candidates for correction by the method outlined. The lexicon was then augmented with additional alternatives for those words which needed correction. We repeated this process with the newly revised dictionary, augmenting it again. We then reran the time-align and tuning process to determine if the newer pronunciations would score better than the first. Finally, we ran the regular Sphinx-II decoder on a random sample of Wall Street Journal training data with the original lexicon and the tuned lexicon, and did the same but only for utterances that included the target words and only for female speakers. Thus, this data had not been seen by the models used for tuning, since we used only male data and models for tuning. We also ran a similar series of tests on dictated physical examination data collected by Berdy Medical Systems for the development of a speech recognition and understanding front-end for computerized patient record systems.

4. RESULTS

Of 13238 words in the WSJ lexicon and training data, the lexical tuning process found 379 candidates for correction. Of these an average of 1.2 out of 7.8 phones per word had a z-score of less than -2.0. Upon manual inspection, 247 candidates required that 250 new pronunciations be added as alternatives to those already in the lexicon. Of these 23 involved serious errors in transcription. For example:

"Chagall" [ch AE G AX L]
vs. [SH AX G AO L]

"Keynesians" [K ey N iy S IY AX N Z]
vs. [K EY N Z IX AX N Z]
or [K IY N Z IX AX N Z]

In these and other examples, lower case phones are candidates identified by the lexical tuning process. Six involved adding parts of speech whose pronunciations differed but were not in the lexicon. For example:

"contest" [K aa N T EH S TD]
vs. [K AX N T EH S TD]

The remainder involved minor changes to account for fluent speech, different dialects, phonetic variants, variance in phonetic transcription protocols, or other minor transcription errors, including many examples of de-emphasized vowels and deletions. For example:

"culinary" [K y UW L IX N EH R IY]
vs. [K AX L IX N EH R IY]
vs. [K UW L IX N EH R IY]

Several of these latter reflected our hypotheses about differences between acoustic classifications of sounds and traditional linguistic classifications. For example, [T R] in "trap" is the traditional broad interpretation, but the /t/ is often palatalized in the rhotic context. This should be handled by context-dependent phones. Nonetheless, we added [CH R AE PD] to account for the low-scoring initial /tr/.

Manual corrections were conservative. Typically only one or two phones, and only phones in the vicinity of the candidate phone identified by the lexical tuning program were changed from the original transcription. For example, the correction we entered for the incorrect pronunciation of "Chagall" [ch AE G AX L] modified only the candidate /ch/ phone, e.g. [SH AE G AX L], not the other obviously incorrect vowels.

Inspecting the results, a number of correction patterns become apparent. Some of these are summarized in Table 1. Column one is the word label, column two is the original phonetic transcription, and column three is the corrected pronunciation. In column two, lower case phones are candidates identified by the lexical tuning program. In column three, lower case phones indicate those which were changed when correcting the original. Deletions are indicated by an underscore for emphasis. Some of the more prevalent change patterns were vowel neutralizations (e.g., AA → AX, EH → IX, AO R → AXR), neutral vowel height changes, especially AX → IX preceding alveolars, deleted schwa, and corrected vowel shifts missing from transcriptions of morphological variants. Despite these patterns, we did not attempt to generalize the changes throughout the entire dictionary. Changes were limited to individual candidate lexical entries. Context-sensitive generalization probably would multiply the effectiveness of this technique considerably [5].

Once these additions were made to the dictionary, the time-align decoder was run and lexical tuning statistics were compiled again. Now there were 279 candidates, of which 41 were new and 238 overlapped with pass one candidates. However, only 125 of the words we adjusted after the first pass still had low scoring phones, and many phones scoring low during the first pass no longer did after the corrections, despite the other residuals. The 41 new candidates were revealed because phone score distributions changed as result of the first pass of corrections. We then added additional variants for some words that failed the first pass and for several of the 41

new candidates. The augmented dictionary contained 314 new pronunciation alternatives involving 271 words, 1.8% of the lexicon.

Once dictionary modifications were complete, we ran the forced alignment process on the training set once again to recompute triphone confidence scores. The frequency with which a corrected pronunciation is chosen instead of the original low-scoring pronunciation is a preliminary measure of the correction's effectiveness. Of 314 corrected pronunciations, 190 (60.5%), representing 184 words, were chosen over the original pronunciations for at least one instance of the target word. Fifty-two of the corrected pronunciations were preferred in all instances, effectively replacing the original pronunciation. However, most corrections did not supplant the original completely. Among all the instances of the 184 words for which the correction was at least partially effective, the corrected pronunciation was preferred an average of 60.3% of the time.

When we ran the recognizer on a random selection of training data, we achieved a 1.4% word error rate reduction from 7.79% to 7.68% commensurate with the number of words changed in the lexicon.

The lexical tuning statistics were gathered from male speaker training data. We tested the new dictionary on female training data. Among utterances that included at least one of the words whose lexical entry was augmented, we achieved an 11.7% reduction in word error rate from 7.55% to 6.67%. Among the 271 words with corrected pronunciations, 148 words were recognized accurately with the original dictionary, 39 words were misrecognized at least once with the original dictionary but were perfectly recognized with the new dictionary, 17 words showed some error reduction, 35 words showed no improvement in recognition, and 32 words did not occur in the test set. No words showed an error increase.

Forced alignment rescoring was a partial predictor of improved recognition. Of words whose corrected pronunciations were preferred over the original transcriptions, 27.2% showed improved recognition in the female test set. Recognition errors did not decrease for all such entries because many (60%) were correctly recognized despite low scoring phones, and because of the effect of acoustic model or language model deficiencies. Only 6.9% of corrected pronunciations for which forced alignment showed no preference resulted in fewer recognition errors. This latter effect is probably due to the effect of variants present in the test set but not in the training set.

The effectiveness of various patterns of transcription corrections is illustrated in Table 1. The fourth column reports the forced alignment preference for the corrected pronunciation. The denominator reports the number of instances of the word in the training set. The numerator indicates the frequency with which the corrected pronunciation was chosen by the time-alignment process. The last column reports the number of errors observed in the test set with the original and corrected dictionaries, respectively. Thus, the new pronunciation for "cooperative" was chosen by forced alignment for all seven instances of the word in the male training set, and the error count in the female test set was reduced from 2 to 1. Examples in Table 1 are limited to those corrections judged potentially effective by the forced alignment rescoring. Only a representative sample is listed due to space limitations.

Among the types of changes made, several patterns were effective in reducing errors, though highly context-sensitive or lexically specific. For example, neutralizing unstressed /AA/ to /AX/ or /AA R/ to /AXR/ substantially reduced errors for "comparable", "composite", "participants" and other words, neutralizing the unstressed /OW/ in "micro" improved recognition of "Microsoft" and "microcomputer", but not "microchip", but neutralizing unstressed /AO R/ to /AXR/ had mixed results. Several changes did not seem effective or necessary based on test set error rates. For example, there were no errors before or after corrections when (1) changing alveolar stops to flaps between /N/ and neutral vowels, (2) inserting /T/ between /N/ and /S/, (3) neutralizing tense vowels like /AE/ to /EH/, (4) palatalizing word-initial /T/ before /R/, or (5) deleting schwa in some contexts.

Finally, we applied this process to a corpus of physical examination records dictated by physicians and medical students. The dictionary of 10466 words was tuned using an acoustic training corpus of 13700 utterances (about 1/3 female, 2/3 male), resulting in added pronunciations for 126 words, 1.2% of the dictionary. It was tested on a dataset of 1423 female and 897 male utterances. Of the 126 tuned words, only 20 were represented in the test set. Nonetheless, we observed a reduction in word error rate of 1.64% for the female test set and 2.17% for the male test set. Eight of the 20 tuned words in the test set showed a decrease in errors; none showed an increase.

5. DISCUSSION

We have demonstrated the effectiveness of a practical lexical tuning method on the Wall Street Journal training corpus. Preliminary results on a computerized medical patient record lexicon are promising. Thus far we have corrected low-scoring pronunciations by hand. We are in the process of automating the procedure.

The overall effectiveness of the method might be improved by generalizing correction patterns to similar

phonetic contexts throughout the dictionary. Without such generalization, the method is limited by available data. Furthermore, our experience thus far suggests that without multiple exemplars of each word, lexical tuning results merely reflect variance among instances of each word. Its effectiveness might also be improved by subsequent acoustic model retraining or tuning. We have not yet evaluated these enhancements.

ACKNOWLEDGMENTS

This work was supported in part by ATP Cooperative Agreement 70NANB5H1184 from the U.S. National Institute of Standards and Technology.

REFERENCES

[1] Cohen, M., *Phonological structures for speech recognition.* Ph.D. Dissertation, Dept. of EE and CS, University of California, Berkeley, CA, 1989.

[2] Wooters, C. and Stolcke, A., Multiple-pronunciation lexical modeling in a speaker independent speech understanding system. ICSLP-1994, pp. 1363-1366.

[3] Schmid, P., Cole, R., and Fanty, M., Automatically generated word pronunciations from phoneme classifier output. ICASSP-1993, pp. II-223-226.

[4] Ravishankar, M.K., *Efficient algorithms for speech recognition*. Ph.D. Dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, May 1996.

[5] Eskenazi, M. & Ravishankar, M.K., personal communication.

Table 1. Some Examples of Patterns of Lexical Tuning Corrections and Their Effectiveness

Word	Original pronunciation	Corrected pronunciation	Align-ments	Recog. Errors
Contracted possessive				
AKERS'S	EY K AXR Z ix z	EY K AXR Z _	3/4	3/1
Corrected morphological vowel shift				
COOPERATIVE	K OW AA P AXR ey DX IX V	K OW AA P r ax DX IX V	7/7	2/1
Deleted syllable-final /n/				
CONCENTRATING	K AA N S AX n T R EY DX IX NG	K AA N S AX _ T R EY DX IX NG	5/8	1/0
Flap after /n/				
POINTING	P OY N t IX NG	P OY N dx IX NG	5/6	0/0
Neutral vowel height				
FOUNTAIN	F AW N T AX n	F AW N T ix N	1/2	1/1
GALVESTON	G AE L V ax S T AX N	G AE L V ix S T AX N	3/5	0/0
Neutralization				
AMBASSADOR	ae M B AE S AX DX axr	eh M B AE S AX DX AXR	4/5	0/0
COMPOSITE	K aa M P AA Z AX TD	K ax M P AA Z AX TD	33/49	13/2
ENGAGES	eh N G EY JH IX Z	ix N G EY JH IX Z	4/4	0/0
FORESEES	F ao R S IY Z	F axr S IY Z	1/2	0/0
INDONESIA	IH N D ow N IY ZH AX	IH N D ax N IY ZH AX	10/10	5/0
MICROCHIP	M AY K R ow CH IH PD	M AY K R ax CH IH PD	2/3	0/0
MICROSOFT	M AY K R ow S AO F TD	M AY K R ax S AO F TD	6/6	3/1
MIDLAND	M IH D L ae N DD	M IH D L ix N DD	6/9	1/0
PARTICIPANTS	P aa R T IH S AX P AX N TS	P axr T IH S AX P AX N TS	19/26	4/1
PRECISION	P R iy S IH ZH AX N	P R ix S IH ZH AX N	9/9	1/0
PREMIER	P r EH M IH R, P R iy M IH R	P R ix M IH R	7/10	0/0
PREPARE	P r IY P EH R	P R ix P EH R	11/11	2/0
PROCUREMENT	P R ow K Y UH R M AX N TD	P R ax K Y UH R M AX N TD	13/21	3/0
REMINDER	R iy M AY N D AXR	R ix M AY N D AXR	11/11	1/0
Neutralization and tapping exceptions				
DEVALUED	D ix V AE L Y UW DD	D iy V AE L Y UW DD	3/3	1/0
MITTERRAND	M iy dx AXR AE N DD	M IY t AXR AA N DD	10/10	4/2
Part of speech, morphological, or other phonetic variant				
CONTROVERSIAL	K AA N T R AX V ER SH ax L	K AA N T R AX V ER SH y AX L	8/13	2/2
INTIMATE	IH N T AX m ey TD	IH N T AX M ix TD	2/4	0/0
PRESTIGIOUS	P AXR S T IY jh AX S	P r ix S T IH JH AX S	4/8	0/0
RECORDS	R AX K ao R D Z	R AX K axr D Z	19/23	10/3