SPEAKING MODE DEPENDENT PRONUNCIATION MODELING IN LARGE VOCABULARY CONVERSATIONAL SPEECH RECOGNITION

Michael Finke and Alex Waibel
finkem@cs.cmu.edu, ahw@cs.cmu.edu
Interactive Systems Laboratories Carnegie Mellon University (USA)

ABSTRACT

In spontaneous conversational speech there is a large amount of variability due to accents, speaking styles and speaking rates (also known as the speaking mode) [3]. Because current recognition systems usually use only a relatively small number of pronunciation variants for the words in their dictionaries, the amount of variability that can be modeled is limited. Increasing the number of variants per dictionary entry is the obvious solution. Unfortunately, this also means increasing the confusability between the dictionary entries, and thus often leads to an actual performance decrease. In this paper we present a framework for speaking mode dependent pronunciation modeling. The probability of encountering pronunciation variants is defined to be a function of the speaking style. The probability function is learned through decision trees from rule based generated pronunciation variants as observed on the Switchboard corpus. The framework is successfully applied to increase the performance of our state-of-the-art Janus Recognition Toolkit Switchboard recognizer significantly.

1. INTRODUCTION

Spontaneous, conversational speech tends to be much more variable than the careful read speech that much of speech recognition work has focused on in the past. Not surprisingly the recognition accuracy is much lower on spontaneous speech. Pronunciation differences, in particular, represent one important source of variability that is not well accounted for by current recognition systems. For example, the word "BECAUSE" might be pronounced with a full or a reduced vowel in the initial syllable (IY vs. AX, respectively), or the whole initial syllable might be dropped (as in "CUZ"). Increasing the allowed pronunciation variability of words is needed to handle the reduction phenomena that seem to be a cause of many errors in conversational speech. Unfortunately, as many researchers have noticed, simply increasing the allowable set of pronunciations in all contexts often does not help. In fact, it may even hurt performance, since the gain of including more pronunciations may be offset by a loss due to increased confusability.

If it is the case that pronunciation changes are systematic, then the set of allowable pronunciations (or their likelihoods) can be varied dynamically thereby reducing the added confusability. Thus, the goal of the work presented in this paper is to develop a method for allowing pronunciation variations depending on a hidden speaking mode [3]. The speaking "mode" would vary within and across utterances and would reflect speaking "style", e.g. indicating the likelihood of reduced or sloppy speech vs. clear vs. exaggerated speech. By changing the allowed pronunciations as a function of the speaking mode, we can account for systematic variability without increasing the confusability associated with a static model.

In this paper we present a new approach for modeling pronunciation variation dependent on the speaking mode as implemented in the Janus Recognition Toolkit (JRTk). We present first results based on the JRTk Switchboard Large Vocabulary Conversational Speech Recognizer [1].

2. EXPLORATION AND PREDICTION

Unfortunately, in Switchboard as well as CallHome, the two databases of recorded conversational telephone speech we are considering here in this paper, there is only very limited information on possible pronunciation variations available in the transcriptions:

> Do not try to imitate pronunciation; use a dictionary form: "no" will do for "naw," "nah," etc., "oh" for "aw,"; "going to" (not gonna or goin to); "you all" rather than "y'all"; "kind of" instead of "kinda"; etc. ... Contractions are allowed, but be conservative. For example, contraction of "is" (it's a boy, running's fun) is common and standard, but there'll (there will) be forms that're (that are) better left uncontracted. It is always permitted to spell out forms in full, even if the pronunciation suggests the contracted form.
>
> *Switchboard Transcription Manual*

That means, that in order to predict pronunciation phenomena depending on a speaking mode, we first have to come up with a model of which pronunciation variants we expect to find in the database (*Exploration*). The corpora transcriptions will only be of limited use since variations in pronunciation within a word are not transcribed at all (stress patterns, reduction of unstressed vowels or syllables) and cross word pronunciation phenomena like contractions (e.g. "he is" goes to "he's") and reductions ("going to" becomes "gonna") are transcribed in an inconsistent way.

In this paper we develop a probabilistic model based on context dependent phonetic rewrite rules to come up with a list of possible pronunciations for a word or a sequence of words. The idea is then to automatically retranscribe the corpus based on the variants allowed in order to train a model of how likely which form of variation is and of what the likelihood of a variant being observed in a certain context (acoustic, word, speaking mode or dialogue) is (*Prediction*).

3. MULTIWORDS

Crossword pronunciation phenomena like contractions and reductions are especially hard to handle in state-of-the-art speech recognition engines. The unit of training and recognition in speech recognizers are typically words. Even though allophonic modeling takes the neighbouring phones into account, there are no means so far that allow for reduction/rewriting of phones in a word depending on word context. Ignoring the word neighbours and still allowing for all sorts of phonetic reduction would result in a long list of confusion pairs of very frequent words. Consider of example word sequences like "KIND OF" and "SORT OF" which are often reduced to "KINDA" and "SORTA". If, in order to capture this reduction of "OF", we would introduce the pronunciation variant "OF(A)" transcribed with the unstressed vowel AX the confusability in the dictionary would increase significantly.
In order to model crossword pronunciation phenomena at least for very frequent sequences of words, we picked a list of 205 so-called multiwords and added them to the dictionary.

⋮ ⋮ ⋮ ⋮ ⋮

DID˙HE DID˙YOU DOES˙THAT DOES˙THIS DON'T˙HAVE DON'T˙WANT DO˙IT DO˙WE DO˙YOU DO˙YOU˙HAVE FIND˙OUT FOR˙A FOR˙AN FOR˙THE GET˙A GET˙OUT˙OF GIVE˙ME GOING˙TO GOING˙TO˙BE GOING˙TO˙HAVE
GOT˙A GOT˙TO GOT˙YOU GOT˙YOUR GO˙TO

⋮ ⋮ ⋮ ⋮ ⋮

KIND˙OF LET˙ME LIKE˙A LOOK˙AT LOT˙OF

⋮ ⋮ ⋮ ⋮ ⋮

The criterion for combining words to multiwords was twofold: 1) mutual information between words, and 2) reduction in bigram perplexity (considering the multiword as a new language model token). It turns out that most of the multiwords consist of at least one of the short function words A, AND, AT, IT, OF, or TO. The initial phonetic transcription of multiwords in the dictionary consisted of the concatenation of the transcriptions of the multiword's components.
Having multiwords in the dictionary the question is how to treat these words in the decoding pass. We could either train our language model on a text file where sequences of words are replaced by multiwords or split multiwords when it comes to compute the LM probability for a given sequence of words. Table 1 lists test results for an unadapted test of last year's NIST-Hub-5 evaluation set using the JRTk Switchboard recognizer. Based on these numbers not modeling multiwords in the language model yields significantly better performance.

Condition	SWB WER	CH WER
LM without Multiwords	34.1%	47.2%
LM with Multiword tokens	34.4%	48.1%

Table 1. Multiwords in LM.

4. PRONUNCIATION MODELING IN JRTK

The next step is to expand the recognition dictionary by applying a set of phonological rules in order to generate a variety of pronunciation variants. A sample of these rules is given in Table 2. These rules are applied to all words with the matching context, so they can generalize to new or unobserved words.

1	[AX IX] N → (E)N
2	[AX IX] M → (E)M
3	[AX IX] L → (E)L
4	[AX IX] R → AXR
5	[T D] → DX / [+VOWEL] __ [AX IX AXR]
6	[T D] R → DX
7	L → 0 / __ Y [AX IX AXR]
8	IY → Y / __ [AX IX AXR]
9	NG → N
10	HH → 0 / WB __
11	W → 0 / WB __
12	DH → 0 / WB __
13	[T D] → 0 / [+VOWEL] __ [TH DH]
14	[T D] → 0 / [+CONS +CONTINUANT] __ WB
15	R AX → ER / [-WB] __ [-WB]
16	T → 0 / [M N NG] __ [AX IX AXR]
17	BECAUSE → K [AH AO] Z
18	GOING TO → G AH N AX
19	WANT TO → W AH N AX
20	YOU KNOW → Y AX N OW
21	DO YOU → D Y UW

Table 2. Pronunciation transformation rules used in JRTk.

4.1. Flexible Transcription Alignment (FTA)

Once an expanded dictionary is created, forced alignment (viterbi) is used to determine which pronunciation is associated with each word token in the training corpus.
There are several sources of errors in the transcriptions of the Switchboard corpus

- Pronunciation: As discussed above there is no consistent way of handling contractions in those transcripts. That means that even when the transcription says "THERE WILL" it might very well be "THERE'LL" instead.
- Segmentation: In Switchboard utterance boundaries are not well defined. It turned out that a lot utterances were split incorrectly into utterances such that words at the beginning or end of an utterance were either only partially existent or not there at all.
- Incorrect Transcriptions.

In order to train our speech recognizer based on such unreliable transcriptions we implemented a Flexible Transcription Alignment (FTA) procedure in JRTk [1]. Instead of aligning the plain transcription of an utterance we generate a hidden markov model for each utterance that allows for

1. all alternative pronunciations in the dictionary for each word,
2. multiwords as alternative word to the sequence of words they consists of,
3. beginning and ending words of an utterance being optional,
4. optional silence or breathing models between words,
5. optional noise words to start or end an utterance.

The second component of the FTA approach is a label boosting procedure [5, 1]. Instead of relying on a speaker independent system to align the FTA utterance HMM, we adapt the recognizer using maximum likelihood linear regression (MLLR) to derive a speaker dependent recognizer for each speaker. The speaker dependent forced alignment is then used to determine which of the predicted pronunciation effects really occur in the training database.

$(<BREATH>) <NOISE>(BREATH) $ AND $(<SBREATH>) I $ YOU_KNOW $ IT'S $ I GUESS IT'S SO NORMAL TO(2) $(<BREATH>) START TO WONDER $ ABOUT THAT EVEN IF(2) SHE DOESN'T(2) NEED THAT BUT $(<SBREATH>) YOU_KNOW SHE'S KIND_OF(KINDA/1) ASKING(1/9/9) QUESTIONS ABOUT WHAT(2) $(<BREATH>) WELL WHAT'S GOING_TO(GONNA/1) HAPPEN THIS CAN'T LAST FOREVER(1/4,18,20/18,20) AND $(<SBREATH>) <NOISE>(THROAT)

Table 3. FTA transcript of a Switchboard utterance; parentheses mark pronunciation variants and $ is the silence word.

Table 3 shows the alignment for a SWB utterance. The underlined words were part of the original SWB transcription. Parentheses mark pronunciation variants with the rule numbers that they were derived from attached. In this sample utterance we observe among other things, that the GOING TO goes to GONNA rule was applied, that the ending NG in the word ASKING is reduced to N (rule 9) and that KIND OF surfaced as KINDA.

4.2. Estimation of Prior Rule Probabilities

The number of occurrences of each pronunciation rule can be determined from the number of occurrences of each pronunciation, since a record of the rules is maintained in the dictionary. From this information, the distribution of the different rules can be calculated. Transformation rule probabilities $p(r)$ are estimated from the relative frequencies of each rule r. Since the set of rules was chosen so that they are applied frequently, these estimates are quite robust.

It is possible to improve this prior probability estimates by taking phonetic context and information about the word (unigram count, content vs. function word) in the transformation rules into account. To this end, decision trees were grown to predict $p(r|w)$. For example, rule 10 (HH $\rightarrow$ 0 / WB _), which has the biggest gain in classification accuracy, is associated with a tree with three leaves:

```
R'SYLLABIC = f:  0.8255 0.1745
R'SYLLABIC = t:
—   R'CENTRAL-VOW = f:  0.1509 0.8491
—   R'CENTRAL-VOW = t:  0.5863 0.4137
```

To evaluate the goodness of the trees, we can compare classification performance of $p(r|w)$ to using $p(r)$ only. This comparison is shown in Figure 1. The results show that improved results are obtained for three of the rules.

5. INTRODUCING MODE DEPENDENCIES

It turned out to be much more effective to take other speaking mode related features into account when predicting the probabilities of pronunciation variants. Various measures of speaking rate (word/phone rate etc.), deviation from mean word/phone duration of w, F0, etc., showed to be very good predictors of the probability distributions of the pronunciation variants.

In order to learn the probability of applying a rule given phonetic and linguistic context w and speaking mode related context information m, decision trees were grown to predict $p(r|w,m)$. Note that this use of decision trees

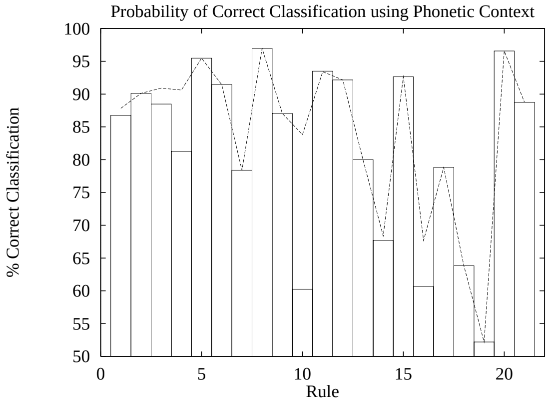

Figure 1. Classification performance of Janus pronunciation rules using decision tree prediction from word features $p(r|w)$ (dotted lines), compared to chance performance (bars).

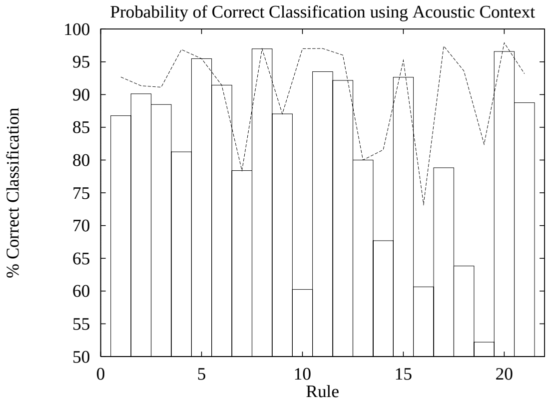

Figure 2. Classification performance of Janus pronunciation rules using decision tree prediction from word and mode features $p(r|w,m)$ (dotted lines), compared to chance performance (bars).

for pronunciation modeling is different from the approach proposed by Riley [4]. The main contribution of our approach is to include speaking mode related information into the procedure of predicting the probabilities. To evaluate the goodness of the resulting trees, we can compare classification performance of $p(r|w,m)$ to using $p(r)$ only as shown in Figure 2. Several rules show major gains in performance with mode conditioning. From informal inspection of the resulting trees, it appeared that relative duration cues were the most important factors.

Mode conditioning leads to a significant gain in classification accuracy for example for rule 18 (GOING TO $\rightarrow$ G AH N AX). The decision tree for this rule mainly consists of speaking rate and word/phone duration related questions:

```
nDur < 2.419:
—   wlen < 1.0129:
—   —   nDur < 1.7819:
—   —   —   wlen < 0.87295:
—   —   —   —   ROWIMD < 0.034522:
—   —   —   —   —   enrateNorm < 1.3588:  0.254 0.746
—   —   —   —   —   enrateNorm >= 1.3588:  0.6667 0.3333
—   —   —   —   ROWIMD >= 0.034522:
—   —   —   —   —   wlen < 0.8383:
—   —   —   —   —   —   enrateNorm < 1.096:
—   —   —   —   —   —   —   ROWIMD < 0.040523:  0.3333 0.6667
—   —   —   —   —   —   —   ROWIMD >= 0.040523:  0.7857 0.2143
—   —   —   —   —   —   enrateNorm >= 1.096:  0.7619 0.2381
—   —   —   —   —   wlen >= 0.8383:  6.25e-09 1
—   —   —   wlen >= 0.87295:  0.8611 0.1389
—   —   nDur >= 1.7819:  0.05172 0.9483
—   wlen >= 1.0129:
```

```
— — sDur < 0.43458:
— — — nDur < 2.0528:
— — — — sDur < 0.23198: 4.167e-09 1
— — — — sDur >= 0.23198: 0.903 0.09697
— — — nDur >= 2.0528: 6.494e-10 1
— — sDur >= 0.43458: 0.9691 0.0309
nDur >= 2.419: 6.596e-11 1
```

6. PROBABILISTIC MODEL

Once rule probabilities are obtained, they can be used to provide the probabilities of each pronunciation variant in the dictionary. Let $\mathbf{r}^+$ be the rules which match with the baseform of w and were applied to derive variant $q_i(w)$. Let $\mathbf{r}^-$ be the rules which match with the baseform but were not used to get $q_i(w)$. The probability of pronunciation variant $q_i(w)$ is given by:

$$P(q_i|w) \quad = \quad \frac{\prod_{\mathbf{r}^+} P(\mathbf{r}^+) \prod_{\mathbf{r}^-} (1 - P(\mathbf{r}^-))}{Z}$$

where Z is a normalization constant.

7. RESULTS

The test set for the evaluation of our mode dependent pronunciation modeling approach consists of the Switchboard and CallHome partitions of the 1996 NIST Hub-5e evaluation set. All test runs used the JRTk Switchboard recognizer. The expanded dictionary that was used in these tests included 1.78 pronunciations variants/word, compared to 1.13 for the baseform dictionary (PronLex). The first list of results in Table 4 is based on a JRTk SWB recognizer whose polyphonic decision trees [2] were still trained on viterbi alignments based on the unexpanded dictionary. We compare a baseline system trained on the base dictionary with an expanded dictionary FTA trained system tested in two different ways: with the base dictionary and with the expanded one. It turns out, that FTA training reduces the word error rate significantly, which means, that we significantly improved the quality of the transcriptions through FTA and pronunciation modeling. Due to the added confusability of the expanded dictionary the test with the large dictionary without any weighting of the variants yields slightly worse results than testing with the baseline dictionary.

Condition	SWB WER	CH WER
Baseline	32.2%	43.7%
FTA training/test w.basedict	30.7%	41.9%
FTA training/test w.exp.dict	31.1%	42.5%

Table 4. Recognition results using flexible transcription alignment training and label boosting. The test using the expanded dictionary was done without weighting the variants.

Adding vowel stress related questions to the phonetic clustering procedure and regrowing the polyphonic decision tree based on FTA labels improved the performance by 2.6% absolute on SWB and 2.2% absolute on CallHome. Table 5 shows results for mode dependent pronunciation weighting. We gain about an additional 2% absolute by weighting the pronunciation based on mode related features.

Condition	SWB WER	CH WER
unweighted	28.7%	38.6%
weighted $p(r\|w)$	27.1%	36.7%
weighted $p(r\|w,m)$	26.7%	36.1%

Table 5. Results using different pronunciation variant weighting schemes.

8. CONCLUSION

We presented an approach to incorporate speaking style related information into the probability estimates for different pronunciation variants. In our approach the "speaking mode" is not explicitly represented as input or state of the recognizer but emerges as the set of questions that optimally select the pronunciation variant based on acoustic features as well as word or phonetic context. Preliminary results show a significant increase in the performance of predicting the correct pronunciation variant as well as major improvements in word accuracy through FTA, label boosting and using a probability weighted pronunciation dictionary within the JRTk Switchboard recognizer. The JRTk recognizer based on speaking mode dependent pronunciation modeling as presented here was one of the two winning systems of the 1997 NIST Hub5-e evaluation and thus proved to be state-of-the-art.

9. ACKNOWLEDGEMENTS

This research was partly funded by grant 413-4001-01IV101S3 from the German Ministry of Science and Technology (BMBF) as a part of the Verbmobil project. The JANUS project was supported in part by the Advanced Research Project Agency and the US Department of Defense. The authors wish to thank Mari Ostendorf for useful discussions and active support.

REFERENCES

[1] Michael Finke, Jürgen Fritsch, Petra Geutner, Klaus Ries, Torsten Zeppenfeld, and Alex Waibel. The JanusRTk Switchboard/CallHome evaluation system. In *Proceedings of LVCSR Hub 5 Workshop*, May 1997.

[2] Michael Finke and Ivica Rogina. Wide Context Acoustic Modeling in Read vs. Spontaneous Speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997. IEEE.

[3] M. Ostendorf, B. Byrne, M. Bacchiani, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A.Waibel, , B. Wheatley, and T. Zeppenfeld. Systematic Variations in Pronunciation via a Language-Dependent Hidden Speaking Mode. In *International Conference on Spoken Language Processing*, Philadelphia, USA, 1996.

[4] M. Riley. A Statistical Model for Generating Pronunciation Networks. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, 1991. IEEE.

[5] Torsten Zeppenfeld, Michael Finke, Klaus Ries, Martin Westphal, and Alex Waibel. Recognition of Conversational Telephone Speech using the JANUS Speech Engine. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997. IEEE.