Evaluation of Pronunciation Variants in the ASR Lexicon for Different Speaking Styles

Ingunn Amdal and Torbjørn Svendsen

Department of Telecommunications
Norwegian University of Science and Technology,
N-7491 Trondheim, Norway
{amdal,torbjorn}@tele.ntnu.no

Abstract

One of the challenges in automatic speech recognition is how to handle pronunciation variation. The main causes for pronunciation variation are the speaker (voice characteristics, accent, non-nativeness etc.) and the speaking style (reading, spontaneous responses, conversation etc.). An ASR system has basically two options for modelling the variation on the word and sub-word level: lexical modelling of the pronunciation variation or adaptation, i.e. re-training of the acoustic models. The answer to the question of which technique to choose, or how to combine them, may depend on the speaking style. We have therefore investigated the effects of using pronunciation variants for recognition of read speech, spontaneous dictation, and non-native speech. The variants in the standard purpose lexicon tested gave modest improvements and best results for read speech, which is the speaking style of the acoustic model training set.

1. Introduction

An important issue in pronunciation modelling is to know what variation is better modelled at the lexical level and what can be handled by the acoustic models, (Strik, 2001). Segmental variation, such as allophonic variation, can be handled by the acoustic models using adaptation or training on the target speech. Other types of variation may be better handled at the lexical level, e.g. insertions, deletions, and variation that is present for a group of speakers (e.g. dialects), or is typical for a speaking style. Lexical modelling accommodates longer contexts than acoustic modelling, permitting modelling of syllables and even entire words or phrases (Jurafsky et al., 2001). Yet, allowing many pronunciation variants in the lexicon may increase the confusability and thereby the error rate.

Adequate handling of varying speaking styles is one of the main challenges for Automatic Speech Recognition (ASR). Expanding from the domain of read speech, ASR systems will encounter more variability in the speech, e.g. more pronunciation variants. To handle more speaking styles the variability in pronunciation must be treated properly, but how we should model the pronunciation variation may depend on the speaking style. In this paper we therefore investigate the effects of using pronunciation variants for recognition of read speech, spontaneous dictation, and non-native speech.

A handcrafted lexicon will generally outperform standard purpose lexica on the task for which it is optimized. However, the production of a manually optimized lexicon is costly and in many cases not feasible. It is therefore interesting to evaluate pronunciation variation issues using only publicly available resources. This is particularly interesting regarding portability issues and for languages where available handcrafted resources are limited.

We have used language resources available through the Linguistic Data Consortium (LDC) or in the public domain both for the pronunciation variants and for building the recognizer, as well as the speech data used. We have investigated the use of pronunciation variants both in the training of the acoustic models and in testing for the different speaking styles. In this way we can compare acoustic modelling and lexical modelling of the variation.

Acoustic model training: We have trained two sets of acoustic models:

1. "Canonical": trained using transcriptions based on a canonical lexicon

2. "Variant": trained using transcriptions based on a lexicon with variants

The "Canonical" set will model all the variation by the acoustic models. The "Variant" set will have less variation in the acoustic models, leaving more to lexical modelling. Both monophone (context-independent) and cross-word triphone (context-dependent) models were trained.

Acoustic model adaptation: Acoustic model adaptation is one way to handle variation; this technique may also depend on speaking style, since the seed models we use for adaptation will fit the speaking styles differently. This is especially true for unsupervised adaptation where we rely on the transcription given by the original acoustic models. The performance of acoustic model adaptation is also dependent on the amount of available data. We have looked at adaptation using both 1 and 20 sentences from each speaker.

Lexical task adaptation: Including pronunciation probabilities is shown to give increased performance in many experiments, e.g. (Wester et al., 2000). One way to derive the probabilities of pronunciation variants is to perform a forced alignment on a development set and use frequency counts to estimate the probability. For the non-native speakers we have an adaptation set available and we have therefore used this speaking style to consider the pronunciation probability effect.

The paper is organized as follows: The language resources used are described in sections 2 and 3. The experimental setup is described in section 4, and the evaluation results are given in section 5. Finally, discussion and conclusions are given in sections 6 and 7.

Speaking style	*Code*	*Grammar and Vocabulary*
Read, native	h1	20k, open (nvp)
Spontaneous dictation, native	s9	20k, open (vp)
Read, non-native	s3	5k, closed
Read, native	h2	5k, closed

Table 1: Speaking styles tested

Number of pronunciations	*Number of words*
6	13
5	4
4	107
3	228
2	2822
1	16826

Table 2: Number of pronunciations per word in the CMU lexicon for the 20k WSJ vocabulary

2. Speaking styles

The task chosen is the Wall Street Journal (WSJ) database available from LDC (The Linguistic Data Consortium, 1993). WSJ consists of several test sets with different speaking styles represented. The two "hub" tests h1 and h2 consists of read speech. In addition we have the "spokes", here we have used the spoke 3 with non-native speech and spoke 9 with spontaneous dictation, see table 1. The h1 and s9 tests both have a 20k open vocabulary. The s9 test has verbal punctuation, which is not the case for h1. This is denoted by (vp) and (nvp) respectively in table 1. For each test we have used the corresponding test specific bigram supplied with the WSJ distribution.

3. Lexica

The CMU lexicon is a popular pronunciation lexicon for US English and available for free (Weide, 1998). Alternate pronunciations are marked so the lexicon can be used both as a "surface" lexicon with pronunciation variants and as a canonical lexicon by removing the alternatives. The basis of the lexicon is 20k words extensively proofed and used by the Carnegie Mellon University (CMU). Additional words and pronunciations are added from several unproofed sources, but only words or pronunciations that are found in two or more sources have been used. In the subset used for the 20k vocabulary, 3174 of the words have pronunciation variants. The maximum number of variants for one word is 6; see table 2. In the subset of the lexicon used for the 5k vocabulary, 1060 words have multiple pronunciation variants. On average there are 1.2 variants per word.

We also did some comparative tests using Pronlex available from LDC (The Linguistic Data Consortium, 1995). This lexicon is also widely used but is not free. On the other hand, it is claimed to be more consistent. We trained only one version using Pronlex with context-dependent models and pronunciation variants. The Pronlex is based on more phones than the CMU lexicon; 42 versus 39. The average number of pronunciation variants per word in Pronlex is 1.1.

4. The Recognizer

We have trained a baseline recognizer using the HTK toolkit (Young et al., 2000), using fairly standard methods (Woodland et al., 1994). The training set consists of read speech from 284 speakers (SI-284), a total of 37500 utterances. We used MFCC feature vectors with 13 elements, including normalized energy, and added the first and second derivatives giving a total of 39 elements. Each feature vector was derived from a speech frame with a Hamming window of length 25 ms and a 10 ms frame rate.

The RM-models supplied with HTK were used as seed models. The canonical and variant models were then trained separately using Baum-Welch reestimation. For the context-dependent models, different decision trees for clustering triphone states were built for the variant and canonically trained HMMs.

4.1. Training of canonical and variant models

When training acoustic models using pronunciation variants, the variants are used to retranscribe the training data with forced alignment. In this way the acoustic models are used to choose the pronunciation variant to use for training. The most common scheme is to retranscribe the training data once using monophone models and then use this transcription for the rest of the training. We did some preliminary experiments with this type of variant trained models using 4 iterations at each mixture increase.

To make sure that the pronunciation variants affect the models we also tried a scheme where we retranscribed the data for every increase of the number of Gaussians in the observation probability mixture. For every level of mixture components we used Baum-Welch reestimation for 4 iterations using the transcription from the previous level. Then the reiterated models were used to retranscribe the data by choosing pronunciation variants. This transcription was then used to iterate 4 more times. The two different retranscription schemes were performed both for the monophone and triphone models.

For the canonically trained HMMs we used 4 iterations for every level of mixture components both for monophone and triphone models.

Retranscribing for every mixture update gave higher log likelihood of the training data after each iteration compared to retranscribing once. The difference was, however, minor and could be a result of the larger number of iterations used in the training. The recognition results showed a small increase in performance for most conditions when retranscribing for every mixture update. The differences between the two retranscription schemes were however not statistically significant. All the results presented in the next section are derived using models with retranscription for every mixture update.

We also compared the log likelihood after each iteration of training for the variant and canonical models, see figure 1. There was hardly any difference between using

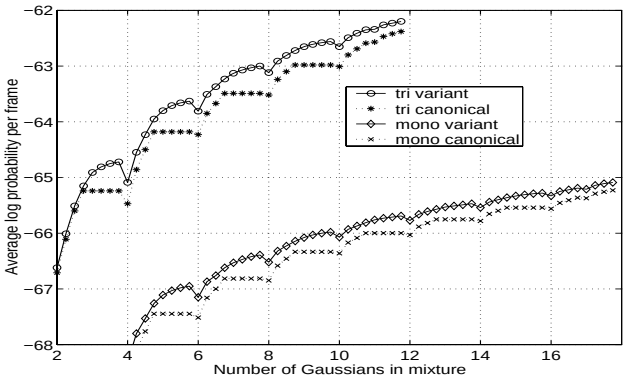

Figure 1: Average log probability per frame after each iteration during training.

transcriptions based on a canonical lexicon or transcriptions with variants. The triphones outperformed the monophones, as expected. For the tests we used 10 Gaussians in the mixture for triphone models. 32 Gaussian monophones did not give substantial improvement over 16 Gaussians for the monophones, so we decided to use 16. We used two silence models with twice as many Gaussians as the phone models.

5. Results

We have used the McNemar test to decide the significance of the differences in word error rate, (Gillick and Cox, 1989). This test takes into account the errors that differ between the systems we compare. The McNemar test requires that errors are independent, which is not the case in our setups since we have applied a bigram. Still, the tests give more information than only word error rate comparisons. We have therefore chosen to use the term "significant" when the McNemar test gives a p-value less than 0.01.

For all the tables in this section the "Training" column shows the lexicon used in training of the acoustic models and the "Test" column shows the lexicon used in test. The last line shows a "mismatch" test where we used variants in acoustic model training but not in the test lexicon.

For the monophone models we observed small improvements by including lexical variants in the test, see table 3. The relative improvements are shown in figure 2. Using the McNemar test the improvements using variants in test only were significant for `h1` and `h2`. The further improvement using variants both in training and test was not significant. For `s9` we had to use variants in both training and test to get significant improvement. For `s3` the improvement was small, there were no significant differences other than the deterioration of the mismatch test. The mismatch test (using variants in training, but not in test) gave significant deterioration compared to using variants in both training and test for all speaking styles.

The improvement seen using variants for the monophone situation was, however, not as uniform for triphones, see table 4 and figure 3. We actually see a deterioration for the two speaking styles `s9` and `s3`. For `h2` we see a deterioration when using variants only in test, but a significant

Training	*Test*	*h1*	*s9*	*s3*	*h2*
Canonical	Canonical	34.7	40.4	27.6	22.2
Canonical	Variant	32.9	39.3	27.2	21.1
Variant	Variant	32.0	38.2	26.7	20.1
Variant	Canonical	35.8	40.8	28.2	22.2

Table 3: WER in [%] for monophone acoustic models

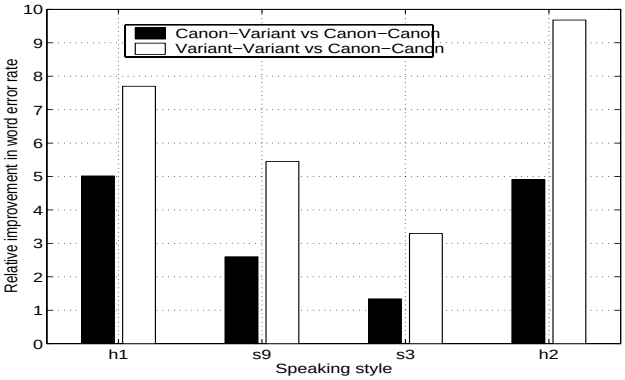

Figure 2: Relative WER improvement from canonical pronunciations in both training and test using monophones

improvement when using the variants in both training and test. There was a significant difference between the mismatch test (using variants in training, but not in test) for all speech types and compared to all other conditions.

In figure 4 the triphone mismatch test is compared to using canonical pronunciations in both training and test and variants in both training and test. The deterioration for variants in acoustic model training and not in the test lexicon was worst for the most controlled situation `h2` (read speech and 5k vocabulary). Both `h1` and `h2` feature the same speaking style as the acoustic model training material. The spontaneous dictation `s9` seems to behave more similarly to `h1` (both 20k vocabulary) than the non-native speech `s3` to `h2` (both 5k vocabulary).

The improvement from context-independent to context-dependent modelling gave the largest difference for `h2`, see figure 5. `s9` behaved again more similar to `h1` than `s3` to `h2`. For `s3` there was no significant change for the canonical setup, and for the variant setup there was actually a significant deterioration! One reason for this is that the variant triphones performed worse than the canonical triphones. The increased modelling capability of the triphones does not help for non-natives on average. As seen in figure 6 the performance for the non-native speakers was very variable.

Training	*Test*	*h1*	*s9*	*s3*	*h2*
Canonical	Canonical	15.9	23.7	27.7	8.0
Canonical	Variant	15.2	23.5	28.4	8.4
Variant	Variant	15.4	24.4	28.9	7.4
Variant	Canonical	20.5	29.4	31.0	11.6

Table 4: WER in [%] for triphone acoustic models

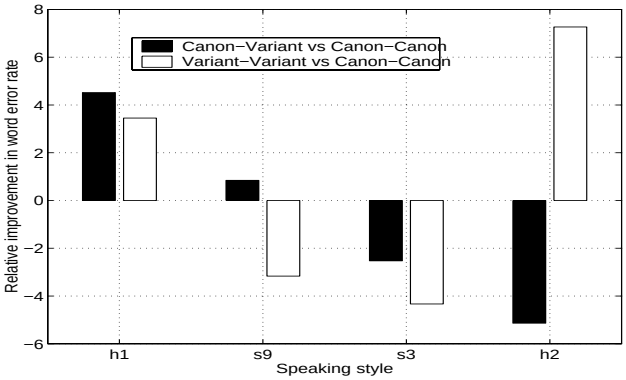

Figure 3: Relative WER improvement from canonical pronunciations in both training and test using triphones

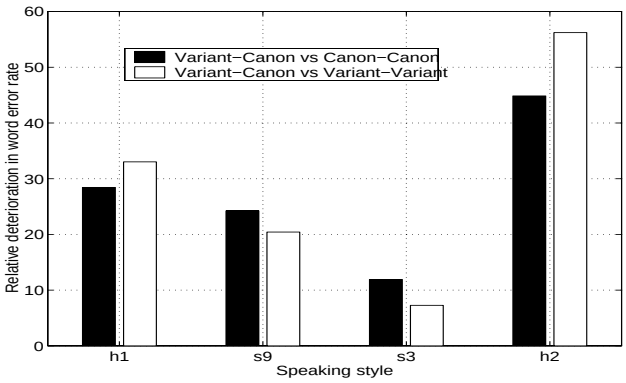

Figure 4: Relative WER deterioration from the two matched conditions to the mismatch condition using triphones

5.1. Comparison between CMU lexicon and Pronlex

Using the Pronlex lexicon gave hardly any difference, see table 5. This indicates that the CMU lexicon is a state-of-the-art non-adapted lexicon. The largest difference was for `h2` with a relative decrease of 6.8%.

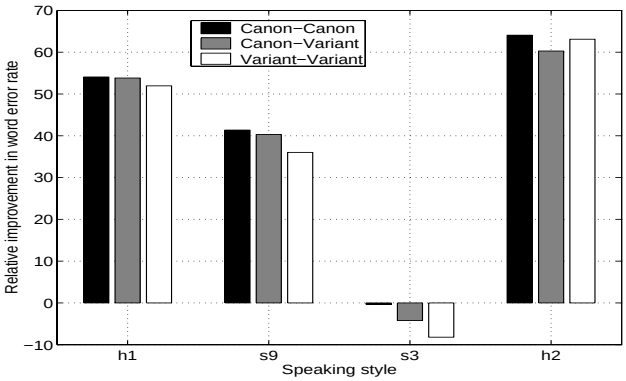

Figure 5: Relative improvement in WER from monophones to triphones

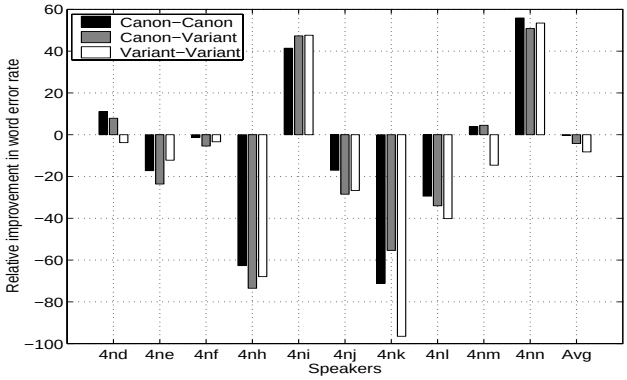

Figure 6: Relative improvement in WER from monophones to triphones per speaker for `s3`

Lexicon	*h1*	*s9*	*s3*	*h2*
CMU	15.4	24.4	28.9	7.4
Pronlex	15.8	24.0	28.5	7.9

Table 5: WER in [%] for variant trained triphone recognizer comparing CMU lexicon and Pronlex

5.2. Adaptation

For acoustic model adaptation we used unsupervised adaptation and the Maximum Likelihood Linear Regression (MLLR) adaptation method, (Leggetter and Woodland, 1995), which is included in the HTK distribution. The adaptation tests were performed using triphone models only, but both for the canonical and variant trained versions. We tried two adaptation schemes:

1. One sentence adaptation where each sentence was used to find one global MLLR transform. The sentence was then re-recognized using this transformation. This will be a reasonable adaptation scheme for telecommunication services where each speaker is active for a short amount of time and we have no information of speaker identity.

2. Incremental adaptation on 20 sentences from each speaker. In this way each sentence from a speaker will update the transform that is used to recognize the next sentence. For this adaptation scheme we used regression classes.

The results in table 6 show that the 20 sentence adaptation as expected gave significant performance gain. Adaptation using variant models gave no or small improvement compared to the canonical models, see figure 7. The largest adaptation gain was for the non-native speakers, whereas the spontaneous dictation gave the smallest gain. The increased performance for the non-native speakers was more uniformly distributed over the speakers compared to using pronunciation variants in the lexicon where the performance varied between speakers, see figure 8. For the `h1` task we see the same increased performance by using variant models over canonical models with adaptation as we saw without adaptation.

HMMs	*h1*	*s9*	*s3*	*h2*
1 sent. canon adap	16.2	23.3	27.3	8.0
20 sent. canon adap	13.4	21.0	20.8	6.6
20 sent. variant adap	13.0	21.3	20.9	6.6

Table 6: WER in [%] for triphone setup and adaptation

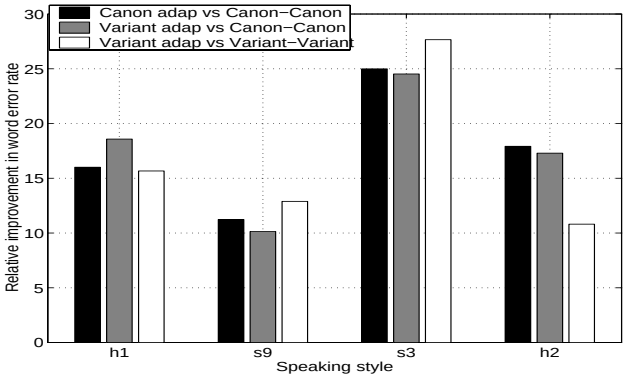

Figure 7: Relative improvement in WER using 20 sentence adaptation on triphone models

As shown in table 6, adaptation using one sentence gave no improvement. Using a regression tree instead of one global transform did not help. One sentence adaptation calls for more sophisticated adaptation methods.

5.3. Error analysis

Even if we got similar recognition rates using canonically and variant trained HMMs, this does not mean that the recognition results are identical. There are both errors corrected and errors introduced, see table 7. For `h1` 25.3% of the errors for the variant system were different from those in the canonical system. This was similar for all speaking styles; 23.9% for `s9`, 29.0% for `s3`, and 27.4% for `h2`. The differences between the canonical system and the system with canonically trained HMMs, but variant lexicon used in test, were not that large: 6.5% for `h1`, 9.8% for `s9`, 12.9% for `s3` and 9.3% for `h2`.

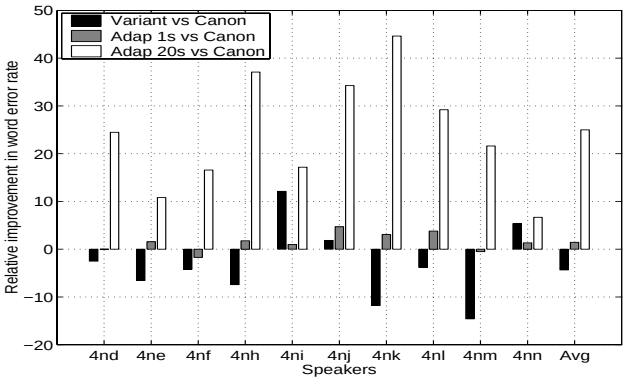

Figure 8: Relative improvement in WER per speaker for `s3` using adaptation on triphone canonical models

Even if the variants in these experiments do not help recognition performance, they make a difference. Our results are similar to the ones shown in (Wester et al., 2000) and shows that improving the recognition performance should be possible by careful selection of which pronunciation variants to include.

5.4. Pronunciation probabilities

Using an adaptation set it is possible to derive speaking style dependent pronunciation probabilities from forced alignment and use these to select which variants to include in an adapted lexicon. This is a simple way of incorporating speaking style dependent lexical adaptation.

As the non-native speakers potentially would gain more from pronunciation modelling we chose this task for some preliminary experiments. In WSJ there is an adaptation set with the same speakers as in the test set available. All speakers read the same 10 sentences, so the vocabulary size of this set is only 349. 256 of these words are present in the 5k test vocabulary. Only 92 of these words had pronunciation variants in the CMU lexicon. This is a small number of words compared to the total number of words in the vocabulary (5000), but we may assume that these are frequent words that are most important to model. The error distribution of function words like “a”, “an”, “and”, “are”, “as” etc. that are present in this 92-word list showed that they are involved in many errors. The pronunciations never selected were left out, and 71 words were left with variants. We did experiments both with pronunciation probabilities added and without, in both cases removing the pronunciations not seen in the adaptation set.

The experiments did not show any improvement, even with different values of the pronunciation scaling factor. In fact, we saw a small deterioration. The reason may be either that this approach calls for larger amounts of adaptation data than we have available, or that the variants present in the CMU lexicon were not representative for the non-native speakers. Collecting sufficient speaking style dependent pronunciations “by hand” is infeasible. There is a need for other methods of deriving and assessing pronunciation variants that are more automatic and more consistent with WER. Data-driven pronunciation variation modelling is one answer, but requires sufficient amounts of representative language resources.

6. Discussion

For the standard purpose lexicon investigated, the CMU lexicon, we could only observe an increase in performance by including pronunciation variants for all speaking styles when using context-independent models. The increased modelling capacity of context-dependent models could apparently handle the observed variation just as well as the variants in the CMU lexicon. We observed, however, that the errors differed: About 20% of the errors were different when using variants compared to using only canonical pronunciations.

Preliminary experiments using forced alignment on an adaptation set to filter out non-useful variants or include pronunciation probabilities did not help. Other language

		Canonical Training + Test	
		Correct	*Incorrect*
Variant Training + Test	Correct	2888	153 (27.9%)
	Incorrect	134 (25.3%)	396

Table 7: Error analysis on word level for `h1`

model issues except these preliminary pronunciation probability experiments were not investigated. It would be interesting to examine how the language model influences the performance for the different speaking styles.

The use of context-dependent models gave a large gain for all speaking styles except non-native speech. This is the same observation as cited in (Van Compernolle, 2001). Triphones trained on native speech are apparently not very good for modelling non-native speech, and more sophisticated methods for this difficult task are necessary. Acoustic model adaptation gave the largest gain for non-native speakers, while the spontaneous dictation gave least improvement.

For pronunciation variation depending on speaking style, the question of lexicon versus acoustic model adaptation has no clear-cut answer. The results presented in this paper show a discouraging performance when lexical variants were included, only small gains. We observe that it seems more important to *not* include variants in the training when we are uncertain which variants will be used in the test. Models trained without variants were able to use a lexicon with variants with equal or better performance. The models trained on variants showed a significant decrease in performance when using a canonical lexicon in test. This is probably because these models are less diffuse and therefore more tailored to the lexicon they are trained on.

On the other hand, the poor gain when using variants may be interpreted as a demand for more care in generation of pronunciation variant candidates and the selection of them. We believe that pronunciation variation modelling is an important factor for improvement of ASR systems. Data-driven variant generation and lexicon optimization using an objective criterion (Holter and Svendsen, 1999), is one such technique.

7. Conclusions and further work

The variants in the standard purpose lexicon tested gave modest improvements. The largest improvements were seen for context-independent models where all speaking styles except non-natives observed a significant improvement. For the context-dependent models the variants did only help for read speech which is the speaking style of the acoustic model training set. For non-native speech we saw no improvement from context-independent to context-dependent modelling.

Even if the error rates were similar for the variant setup compared to the canonical setup, the errors differed. This suggests that there is a potential in selecting variants. To learn more about the contribution of the variants in the lexicon a more thorough error analysis is needed.

Two issues are important in pronunciation modelling: 1) candidate pronunciations, and 2) a way to assess these pronunciation variants. To assess pronunciation variants we need representative data. Methods based on pronunciation rules instead of directly on variants can generalize to pronunciations not present in the training data and will make it possible to assess these unseen pronunciation variants.

8. References

L. Gillick and S. J. Cox. 1989. Some statistical issues in the comparison of speech recognition algorithms. In *Proc. ICASSP-89*, pages 532–535, Glasgow, Scotland.

Trym Holter and Torbjørn Svendsen. 1999. Maximum likelihood modelling of pronunciation variation. *Speech Communication*, 29:177–191.

Dan Jurafsky, Wayne Ward, Zhang Jianping, Keith Herold, Yu Xiuyang, and Zhang Sen. 2001. What kind of pronunciation variation is hard for triphones to model? In *Proc. ICASSP-2001*, Salt Lake City (UT), USA.

C. J. Leggetter and P. C. Woodland. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9:171–185.

Helmer Strik. 2001. Pronunciation adaptation at the lexical level. In *Proc. ISCA ITRW Adaptation methods for speech recognition*, pages 123–130, Sophia-Antipolis, France.

The Linguistic Data Consortium, 1993. *Wall Street Journal speech database (WSJ)*. [online description]. [cited 2002-03-01]. URL: `http://morph.ldc.upenn.edu/Catalog/LDC94S13A.html`.

The Linguistic Data Consortium, 1995. *CALLHOME American English Lexicon (PRONLEX)*. [online description]. [cited 2002-03-01]. URL: `http://morph.ldc.upenn.edu/Catalog/LDC97L20.html`.

Dirk Van Compernolle. 2001. Recognizing speech of goats, wolves, sheep and . . . non-natives. *Speech Communication*, 35:71–79.

Robert L. Weide, 1998. *CMU Pronunciation Dictionary*. [online]. [cited 2002-03-01]. URL: `http://www.speech.cs.cmu.edu/cgi-bin/cmudict/`.

Mirjam Wester, Judith M. Kessens, and Helmer Strik. 2000. Pronunciation variation in ASR: Which variation to model? In *Proc. ICSLP-2000*, Beijing, China.

Phillip C. Woodland, Julian J. Odell, Valtcho Valtchev, and Steve J. Young. 1994. Large vocabulary continuous speech recognition using HTK. In *Proc. ICASSP-94*, pages II:125–128, Adelaide, Australia.

Steve Young, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho Valtchev, and Phil Woodland, 2000. *HTK Version 3.0*. [online description]. [cited 2002-03-01]. URL: `http://htk.eng.cam.ac.uk/`.