ISCA Archive
http://www.isca-speech.org/archive

ITRW on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology (PMLA2002)
Aspen Lodge, Estes Park, Colorado, USA
September 14–15, 2002

UNSUPERVISED PRONUNCIATION ADAPTATION FOR OFF-LINE TRANSCRIPTION OF JAPANESE LECTURE SPEECHES

Daniel Willett, Erik McDermott, Shigeru Katagiri

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto-fu, Japan
{willett,mcd,katagiri}@cslab.kecl.ntt.co.jp

ABSTRACT

Observing that most variations in pronunciation are strongly speaker and speaking style dependent, and that the introduction of pronunciation variants in a speaker-independent recognition system is of limited success, we refrain from applying multiple pronunciation variants in the speaker-independent case and instead introduce pronunciation variants without supervision when specializing the recognizer for a specific speaker. Our approach is to take the decoder's output after a first recognition pass and to realign it allowing several commonly observed pronunciation variations. In a second decoding pass, the pronunciation variations are integrated into the recognizer, weighted using Maximum Likelihood estimates for the pronunciation variants' likelihoods on the realigned output of the first pass. We observe a small but significant improvement in recognition accuracy compared to the first pass output and conclude that the method is helpful in adjusting the pronunciation modeling structure according to speaker, speaking style and speaking rate. A better prior choice of possible pronunciation variations involving deeper phonetic knowledge would be beneficial for further improvements. We also show experimentally that the improvement gained through pronunciation adaptation does not overlap much with the improvement gained by unsupervised adaptation of the acoustic models, but rather that the achieved WER reductions are additive.

1. INTRODUCTION

In recent years, numerous attempts have been made to introduce multiple pronunciations or graph-based probabilistic pronunciation rules into HMM-based speech recognizers [1, 3, 4, 11, 13, 9]. This work has been motivated by the observation that there is a strong mismatch between canonical dictionary pronunciations and the actually observed pronunciations, in particular for spontaneous speech [1]. Unfortunately, incorporating additional pronunciation variants or general pronunciation rules for better modeling of the data in the Maximum Likelihood sense has the effect of increasing confusability between words despite improving training data likelihood [8]. Different heuristic approaches have been proposed to maintain the ability to discriminate words despite of the introduction of multiple pronunciations. Central idea is to introduce only the most valuable alternative pronunciation variants according to some criterion like training data likelihood. What we propose here is to refrain from using pronunciation rules in the speaker-independent recognition case where, due to numerous different speaking styles of different speakers, their introduction decreases discrimination too much, but to introduce them instead when specializing (adapting) the system to a specific speaker.

The target of this study is Japanese speech recognition. Canonical word pronunciation in Japanese is to a large extent determined by word orthography in terms of the Kana symbols of the writing system. (Kana symbols typically represent either a single vowel or a consonant followed by a vowel; short and long vowels are strictly distinguished.) Canonical pronunciation dictionaries for Japanese speech recognition are usually derived through a simple mapping of the Kana-based orthography to the specific vowel or consonant and vowel symbols. This procedure neglects the well-known fact, often referred to as devocalization, that some vowels in some words or phonetic contexts are not articulated in general while others might not be articulated depending on speaking rate, speaking style and the level of politeness that the speaker chooses. Another pattern of variation is for voiced sounds to be articulated unvoiced and vice versa, and that - possibly strongly depending on speaking rate and phonetic context - short vowels might be articulated long and vice versa. In the current Kana-based single pronunciation practice, the hope is that context-dependent hidden Markov models will manage to model these variations sufficiently well.

An obvious way to tackle these observations more directly is to introduce multiple pronunciations or pronunciation rules that allow for such kinds of variations. Imai et al. [4] report considerable word error rate reduction in a speaker-dependent Japanese recognition system, with a large amount of data available to train good speaker-

dependent pronunciation variant probabilities. In a speaker-independent system, however, we found the introduction of pronunciation variation rules to be counter-productive, decreasing recognition accuracy by apparently decreasing discrimination. This led us to the main topic of this study, the use of expanded pronunciation rules only in a second speaker-adapted pass of decoding - with the possibility of implementing the approach via a cheap lattice or N-best list rescoring stage. In order to achieve some generalization from only very limited data, we do not attempt to find word-specific pronunciation variant probabilities, but instead estimate for each physical triphone specific probabilities for the possible variations.

2. BASELINE SYSTEM

We performed experiments using a Japanese recognition system for full unsegmented lecture speeches. The training set consists of 144 full lecture speeches, spoken by 144 different speakers; the test set consists of seven lecture speeches, spoken by seven speakers that are not among the training speakers. The closed vocabulary has a size of around 20k words. The back-off trigram language model applied in decoding was designed using the transcriptions of the training set. The acoustic models are cross-word context-dependent linear no-skip 3-state triphone HMMs clustered via tree-based state clustering. Each of the 2K physical states uses 16 Gaussian mixture components. The speech features used comprise 12 MFCC coefficients and corresponding Δs and energy coefficients. The acoustic models, originally set up on read speech within the IPA project [5], were retrained by means of forward-backward estimation on the lecture speech training set. For a further description of the task see [10].

2.1. Single Pronunciation Setup

The baseline single pronunciation configuration uses only a single pronunciation per word, derived straight from the Kana writing of each word, as it is the standard approach in Japanese speech recognition. Different symbols (and different HMMs) are used for short and long vowels. In the following, the term 'single pronunciation' or 'SP' refers to this baseline system setup, either in the HMM training phase or in the recognition phase; 'SP HMM' refers to the single pronunciation acoustic models.

2.2. Multiple Pronunciation Setup

This configuration uses the following general pronunciation rules to generate alternative pronunciations from the canonical dictionary-based ones described in the previous section:

- skipping of short vowels
- long articulation of short vowels
- short articulation of long vowels
- voiced articulation of an unvoiced consonant
- unvoiced articulation of a voiced consonant

Clearly, these rules are very broad and do not make use of deep phonetic knowledge.

Speaker-independent Maximum Likelihood estimates of the pronunciation variation weights, as well as of the underlying acoustic models, were set up in an iterative procedure using the speaker-independent training data. The iterative procedure consists of consecutive application of forward-backward HMM training and recalculation of the best path for the given utterance transcription allowing the pronunciation variation rules described above. This resembles the supervised approach outlined in [9] of integrating ML estimates of pronunciation variants in automatic speech recognition. Figure 1 shows an excerpt of the resulting recognition network (for the monophone case) after having seen the word 'desu' being pronounced with the voiced $/z/$ instead of the unvoiced $/s/$ in 1 out of 10 times and with some variations observed concerning the final short $/u/$ sound.

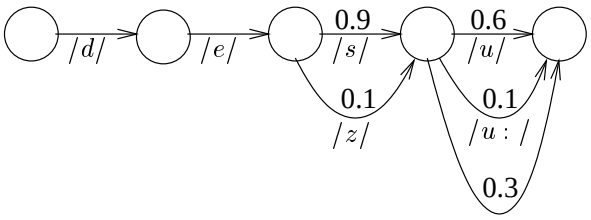

Fig. 1. Pronunciation network for 'desu' in the second pass

As a consequence of this approach, only those pronunciation rules which have been observed at least once in training will be used in recognition, meaning that most rules are disabled in most contexts. The major part of the pronunciation variants seen in training are the skipped vowels $/u/$ and $/i/$, as one would expect from phonological knowledge of Japanese. Long to short variations are also frequently applied, whereas the other variations are rather rare.

In the following, the term 'multiple pronunciations' or 'MP' will refer to the setup that involves the application of pronunciation rules, either in the HMM training phase or in the recognition phase; 'MP HMM' refers to the acoustic models trained in the multiple pronunciation setup.

3. UNSUPERVISED SPEAKER-SPECIFIC WEIGHTING OF PRONUNCIATION RULES

The availability of sufficient speaker-specific transcribed training data allows the estimation of speaker-dependent

pronunciation variants and pronunciation variant probabilities. Speaker-dependent pronunciation variants have successfully been applied in [4] in a supervised way, with a large amount of transcribed speech data from the specific speakers available. In the lecture speech task examined here, we do have a considerable amount of speech data per speaker available. The shortest speech is about 12 minutes long. However, we have no data for each speaker other than the fixed test set, so we can only make use of the data to be transcribed itself in order to derive pronunciation variant probabilities. Hence, the approach evaluated here is an unsupervised estimation of pronunciation rule weights and subsequent use of those weights to aid the recognition of each test utterance. Similar unsupervised model reestimation approaches are common practice for specializing acoustic models to specific speakers and channel conditions [12]; a related approach for unsupervised adaptation of the language model to topic and speaking style is described in [7].

3.1. General approach

Our approach to unsupervised speaker-specific pronunciation weighting is straightforward. We take the output of a first recognition pass, in which speaker-independent multiple pronunciation variants may or may not already have been applied, generate a finite state network for the output model sequence that allows the application of all the pronunciation rules applicable to the output model sequence, and recalculate the best matching sequence in a Viterbi-search through the network. The context-dependent pronunciation variant probabilities found there are then used to weight the pronunciation rules in a second recognition pass.

In the second pass either one of the two sets of baseline hidden Markov models might be used, either the models that have been trained in a consistent approach allowing similar pronunciation variants (MP HMMs), or the models trained without using pronunciation variants (SP HMMS). Figure 2 illustrates this procedure.

3.2. Smoothing and Parameter Tying

Estimating pronunciation variant probabilities on the first pass recognition output raises some questions concerning how to treat unobserved events and events not observed often enough to allow a robust estimation. Our approach is rather simple and certainly leaves some room for improvements. As a first approach to smoothing, we do not estimate word-specific pronunciation variant probabilities, but only triphone specific ones. These are applied for every word that contains the specific triphone. This approach is simplified by the circumstance that our pronunciation variant rules are very simple and all only involve single phoneme variations. As for the treatment of rare events (triphones),

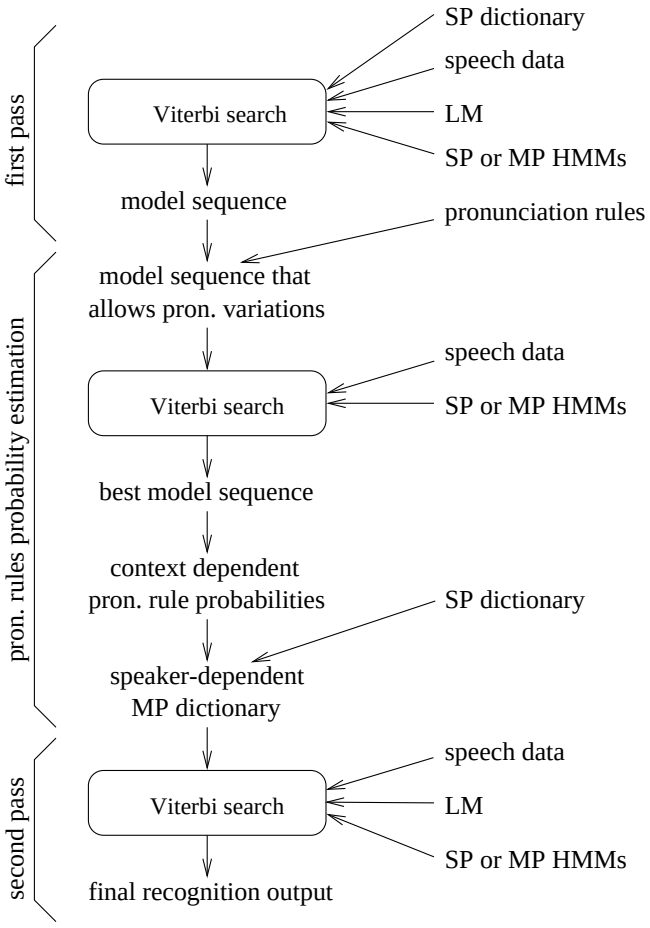

Fig. 2. Multipass decoding process

the probabilities of pronunciation variants of totally unobserved triphones, are estimated by monophone based estimates. No other type of smoothing is performed. Even events (triphones) observed only a few times (possibly only once) are treated in no special way, allowing poorly estimated pronunciation variant probabilities. In this respect, more profound smoothing and tying techniques could yield further improvement.

4. INTEGRATION OF PRONUNCIATION VARIANTS INTO DECODING

When decoding with multiple pronunciation variants per word, two questions arise. One is whether to sum up the pronunciation variants scores or to take only the best matching one into account. For the sake of simplicity in decoding, we chose to use the Viterbi approximation also at the level of the multiple pronunciations per word, which means, that in decoding only the maximum path through the best matching model sequence (in this Viterbi sense) is taken into consideration. The second question concerns the weighting of

the pronunciation variant probabilities. With P_a representing the acoustic model and P_{lm} representing the language model, the decoding task based on the MAP decision rule can be formulated as

$$W^* = \underset{W}{\operatorname{argmax}}\big(P_{lm}(W)^\gamma P_a(X|W)\big). \quad (1)$$

For practical reasons, such as different quality and range of the involved models, it is well known that choosing an appropriate weighting of the involved models through the weighting factor γ is of great importance. Having multiple pronunciations Q that represent a word sequence W and a model P_q which gives us a likelihood of a pronunciation variant for a given word sequence, we can write the decoding problem as the solution of the following equation

$$W^* = \underset{W}{\operatorname{argmax}}\big(P_{lm}(W)^\gamma \max_{Q}\big(P_q(Q|W)^\beta P_a(X|Q)\big)\big). \quad (2)$$

We see that the pronunciation variant model needs some weighting against language model and acoustic models too. In our experiments, we set the pronunciation probability weighting factor β empirically on a few test runs in the cheated experiments (described in 5.2) on just one of the seven test speeches and used this fixed value from there on. Hence, the chosen weighting factor does not represent the optimal value and leaves some room for future improvement. Its estimation on a separate development set would certainly be preferable.

5. EVALUATION

5.1. Baseline

Table 1 shows the baseline recognition performance with the SP and MP acoustic models described in Section 2 with and without allowing pronunciation variants in decoding. The results clearly show that consistency between training and decoding conditions is important and that the best performance is achieved simply by using canonical pronunciations in both training and testing. The pronunciation rules seem to introduce more confusability than they help in terms of discriminating words. However, this is the speaker-independent case and concluding from [4], multiple well-chosen pronunciation rules should be beneficial in a speaker-dependent recognition mode.

	SP in training	MP in training
SP in recognition	36.3%	37.8%
MP in recognition	40.2%	39.8%

Table 1. Baseline performance (word error rate) with and without pronunciation variations

5.2. A Cheating Experiment

This observation motivates the second recognition experiment, in which we set the context-dependent pronunciation rules according to the frequency of appearance in an alignment of the true test utterance transcription. The results from this cheating experiment, summarized in Table 2, show two things. First, they show the upper limit of recognition performance that is possible if the unsupervised setting of pronunciation probabilities estimates the true pronunciation variant likelihoods within the respective test set. Second, the results prove that with an appropriate speaker-dependent weighting of the pronunciation variants, the latter can be usefully applied in the context of this application. Surprisingly, we observe only a minor improvement through the use of the MP HMMs, the hidden Markov model set which was trained allowing multiple pronunciation variants in training.

speech	first pass	second pass SP HMM	second pass MP HMM
a01m0007	34.6%	34.1%	34.1%
a01m0035	42.4%	40.8%	41.2%
a01m0074	28.7%	28.0%	27.9%
a02m0117	36.8%	35.6%	35.4%
a03m0100	36.4%	35.2%	35.3%
a05m0031	35.2%	34.2%	34.0%
a06m0134	40.3%	39.7%	39.3%
mean	**36.3%**	**35.4%**	**35.3%**

Table 2. Word error rate with 'tuned' pronunciation probabilities

5.3. Unsupervised Pronunciation Weighting

In the third recognition experiment, we take the output of the first recognition pass, and train speaker-dependent pronunciation rule probabilities based on this output. Table 3 shows the achieved recognition performance. We gain a small absolute improvement of 0.4% with hardly any difference between the two types of HMM sets.

Tables 4 and 5 list some of the estimated average probabilities for skipping and lengthening a short /u/ at word ends. Table 4 shows the estimates in the supervised 'cheated' mode, that the recognition results in Table 2 are also based on. The strong speaker-dependence of the pronunciation rule application is very clear. Speaker a05m0031 tends to articulate a /u/ at word ends particularly long, whereas speakers a02m0117 and a01m0007 tend to skip it instead.

Table 5 shows the estimates gained without supervision from the realigned recognizer output. Due to the high word error rate, there certainly is a mismatch with the results in

speech	first pass	second pass SP HMM	second pass MP HMM
a01m0007	34.6%	34.4%	34.3%
a01m0035	42.4%	41.6%	41.6%
a01m0074	28.7%	28.4%	28.4%
a02m0117	36.8%	36.2%	36.3%
a03m0100	36.4%	36.2%	36.0%
a05m0031	35.2%	34.8%	34.8%
a06m0134	40.3%	39.9%	39.9%
mean	**36.3%**	**35.9%**	**35.9%**

Table 3. Word error rate with the proposed procedure

speech	devocalization (/u/ at word ends)	short to long (/u/ at word ends)
a01m0007	43%	0%
a01m0035	0%	4%
a01m0074	0%	7%
a02m0117	30%	10%
a03m0100	5%	0%
a05m0031	5%	86%
a06m0134	7%	3%

Table 4. Supervised analysis of the speaker-specific pronunciation rule probabilities (cheated mode)

Table 4. It is obvious, though, that the estimates show the correct tendency.

speech	devocalization (/u/ at word ends)	short to long (/u/ at word ends)
a01m0007	25%	3%
a01m0035	0%	8%
a01m0074	0%	5%
a02m0117	8%	8%
a03m0100	7%	0%
a05m0031	0%	69%
a06m0134	0%	2%

Table 5. Unsupervised analysis of the speaker-specific pronunciation rule probabilities (objective test mode)

This underlines that despite the high word error rate, the proposed unsupervised procedure is well suited for estimating reliable pronunciation variant probabilities.

5.4. Joint Acoustic and Pronunciation Model Adaptation

Table 6 finally lists some preliminary results on combining the proposed unsupervised speaker-dependent pronunciation adaptation with unsupervised adaptation of the acoustic models. It is interesting to see whether the achieved improvements are complementary or whether the unsupervised pronunciation adaptation is covered by acoustic model adaptation. On the one hand, it could be expected that the transformation of the acoustic models can compensate for the speaker-specific pronunciation behavior to some extend. On the other hand, having better fitting pronunciation variants in acoustic model adaptation might also lead to a further model improvement.

Unsupervised HMM adaptation was conducted using HTK's [14] implementation of MLLR adaptation [6], with 250 state clusters and transformation matrices applied on only the Gaussian mean vectors. Pure acoustic adaptation on the single best output in two iterations improves the word error rate (WER) from 36.3% to 31.9% in the single pronunciation setup. Additional iterations have no further impact. The integration of MLLR into the described pronunciation adaptation framework is realized through performing MLLR and pronunciation adaptation in parallel in between the first and the second recognition pass. In this joint adaptation procedure, the second recognition pass uses acoustic models that have been adapted by means of MLLR with respect to the best model sequence, that also the pronunciation variant weight estimation is based upon. Compared to pure MLLR, two iterations of this procedure give an additional absolute improvement of 0.3% to 31.6% WER. This additional improvement is similar to the one observed in Table 3 for pronunciation adaptation without acoustic model compensation. This indicates that the gained improvements are additive.

speech	baseline	MLLR	MLLR+MP
a01m0007	34.6%	32.8%	32.5%
a01m0035	42.4%	40.2%	39.9%
a01m0074	28.7%	24.2%	23.6%
a02m0117	36.8%	29.7%	29.6%
a03m0100	36.4%	32.1%	31.8%
a05m0031	35.2%	27.9%	27.4%
a06m0134	40.3%	36.6%	36.3%
mean	**36.3%**	**31.9%**	**31.6%**

Table 6. Combination with acoustic model compensation

6. CONCLUSION

We have shown that through the unsupervised introduction of pronunciation rules, a slight improvement in recognition accuracy can be gained. The improvement is rather small and possibly does not justify the effort. However, it is a major improvement over the general introduction of pronunciation variants in the speaker-independent case, which is

an interesting observation in itself. It suggests that the introduction of multiple pronunciation variations is promising only for speaker-dependent or speaker-adapted recognition systems. This finding might however differ depending on the language. It should be interpolated from Japanese to other languages with some care. The pronunciation variations considered here were very broad and did not make use of profound phonetic knowledge. In this respect, more carefully chosen pronunciation variability should result in further improvements.

7. ACKNOWLEDGMENT

We thank the Japanese Science and Technology Agency Priority Program, "Spontaneous Speech: Corpus and Processing Technology", for providing speech data and transcriptions, and also the Japanese IPA speech recognition project for providing the initial acoustic models.

8. REFERENCES

[1] M. Eichner, M. Wolff, "Data-Driven Generation of Pronunciation Dictionaries in the German Verbmobil Project - Discussion of Experimental Results", ICASSP'00.

[2] J. E. Fosler-Lussier, "Dynamic Pronunciation Models for Automatic Speech Recognition", PhD Thesis, University of California, Berkley 1999.

[3] I. L. Hetherington, "An Efficient Implementation of Phonological Rules using Finite-State Transducers", Eurospeech'01, pp. 1599-1602.

[4] T. Imai, A. Ando, E. Miyasaka, "A New Method for Automatic Generation of Speaker-Dependent Phonological Rules", ICASSP'95, pp. 864–867.

[5] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, S. Itahashi, "The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus", ICSLP'98, pp. 3261–3264.

[6] C. J. Leggetter, P. C. Woodland: "Speaker Adaptation Using Linear Regression", *Technical Report* CUED/F-INFENG/TR.181, Cambridge University, 1994.

[7] T. R. Niesler, D. Willett, "Unsupervised Language Model Adaptation for Lecture Speech Transcription", ICSLP'02, Denver.

[8] M. Saraclar, S. Khudanpur, "Pronunciation Ambiguity versus Pronunciation Variability in Speech Recognition", ICASSP'00.

[9] F. Schiel, A. Kipp, H. G. Tillmann, "Statistical Modelling of Pronunciation: It's not the model, It's the data", ESCA Workshop on Pronunciation Variation for Automatic Speech Recognition, Kerkrade, 1998, pp. 131–136.

[10] T. Shinozaki, C. Hori, S. Furui, "Toward Automatic Transcription of Spontaneous Speech", Eurospeech'01, pp. 491–494.

[11] T. Sloboda, "Dictionary Learning: Performance Through Consistency", ICASSP'95, pp. 453–456.

[12] F. Wallhoff, D. Willett, G. Rigoll, "Frame Discriminative and Confidence-Driven Adaptation for LVCSR", ICASSP'00.

[13] C. Wooters, A. Stolcke, "Multiple-Pronunciation Lexical Modeling in a Speaker Independent Speech Understanding System", ICSLP'94, pp. 1363–1366.

[14] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev. P. Woodland, "The HTK Book, Ver. 3.1", Cambridge University Engineering Department, 2001.