DICTIONARY LEARNING FOR SPONTANEOUS SPEECH RECOGNITION

Tilo Sloboda, Alex Waibel

sloboda@ira.uka.de,waibel@ira.uka.de
Interactive Systems Laboratories
University of Karlsruhe — Karlsruhe, Germany
Carnegie Mellon University — Pittsburgh, USA

ABSTRACT

Spontaneous speech adds a variety of phenomena to a speech recognition task: false starts, human and nonhuman noises, new words, and alternative pronunciations. All of these phenomena have to be tackled when adapting a speech recognition system for spontaneous speech. In this paper we will focus on how to automatically expand and adapt phonetic dictionaries for spontaneous speech recognition. Especially for spontaneous speech it is important to choose the pronunciations of a word according to the frequency in which they appear in the database rather than the "correct" pronunciation as might be found in a lexicon. Therefore, we proposed a data-driven approach to add new pronunciations to a given phonetic dictionary [1] in a way that they model the given occurrences of words in the database. We will show how this algorithm can be extended to produce alternative pronunciations for word tuples and frequently misrecognized words. We will also discuss how further knowledge can be incorporated into the phoneme recognizer in a way that it learns to generalize from pronunciations which were found previously. The experiments have been performed on the German Spontaneous Scheduling Task (GSST), using the speech recognition engine of JANUS 2, the spontaneous speech-to-speech translation system of the Interactive Systems Laboratories at Carnegie Mellon and Karlsruhe University [2, 3].

1. INTRODUCTION

The phonetic dictionary is one of the main knowledge-sources for a speech recognizer, to lead it to valid hypotheses in the recognition process. Still it is often regarded as being less important as acoustic or language modeling.

In continuous speech recognizers researchers often use the "correct" pronunciation of a word, as it can be found in a lexicon. But this "correct" pronunciation does not have to be the most frequent variant for a given task (especially in spontaneous speech), and does not necessarily yield the best recognition performance given the current acoustic modeling. If the phonetic transcriptions in the dictionary do not match the actual occurrences in the database, the phonetic units will be contaminated during the training with inadequate acoustics, which will degrade the overall performance.

State-of-the-art speech recognition systems start to put more and more effort into creating adequate dictionaries with alternative pronunciations and word contractions, which can also model interword effects such as coarticulation between words (e.g. "gonna" as contraction of "going to") .

As we want to increase the overall performance of the speech recognizer, we are especially interested in the most common pronunciations for the given task, in a better modeling of frequently misrecognized words and strong dialectic variations of word sequences. We will show how our algorithm can learn pronunciations for word tuples and therefore learn interword effects such as coarticulation between words and dialectic variations of words and word sequences.

2. DICTIONARY LEARNING

Modifying dictionaries is usually done either by hand or by applying phonological rules (e.g. [5, 6]) to a given dictionary. Hand tuning and modifying the dictionary requires an expert. It is time consuming and labor intensive, especially if a lot of new words need to be added, e.g. when the task is still growing, or the system is adapted to a new task.

Adding dictionary entries by hand usually focuses on single occurrences of a word and does not have the improvement of the overall recognition performance as an objective function. Furthermore, it is error prone – all the following errors can be introduced when modifying phonetic dictionaries by hand:

- with increasing number of basic phonetic units (usually between 40 and 100) and number of entries in the dictionary, it gets more and more difficult to use the phonetic units consistently across dictionary entries.
- experts tend to use the "correct" phonetic transcription of a word – this is not necessarily the most frequent or even the most likely transcription for a given task.
- actual pronunciations can be very different from the "correct" pronunciation. In spontaneous speech and in dialects a lot of alternative pronunciations are used which are not always easy to predict. The pronunciation

of foreign words and names is also a good example for this (e.g. Goražde, München, Arkansas, Woszczyna).

- as it is hard to say which variants are statistically relevant for a given task, the maintainer of the dictionary can easily miss relevant forms.

If phonological rules are used to derive pronunciation variants, the number of rules can vary between several dozens and more than thousand. Using only a few rules does not necessarily cover all spontaneous effects, using too many rules on the other hand results in too many possible variants. Even applying a few rules to a dictionary increases the number of pronunciations (and therefore increase the computational cost) significantly. Expert knowledge is needed to restrict the application of rules, otherwise overgeneralization of rules can lead to bogus variants. Finally it is not guaranteed that all common variations of a word which appear in a spontaneously spoken corpus are actually modeled by a given set of rules.

Therefore, we propose a data-driven approach to improve existing dictionaries and automatically add new words and variants whenever needed. This algorithm should:

- use a performance driven optimization of the phonetic entries in the dictionary rather than a "canonical" form of a word.
- use the underlying phonetic modeling to generate accurate and consistent entries in the phonetic dictionary.
- generate pronunciation variants only if they are statistically relevant.
- lead to a lower phoneme confusability after retraining.
- lead to a higher overall recognition performance

We give an outline of an algorithm for Dictionary Learning which aims at optimizing the dictionary for retraining, so that contaminated phonetic units will get more accurate training.

In our first experiments we show that even using a simple algorithm to extract candidates for phonetic variants yields a significant increase in recognition performance. We also show experiments of modeling word tuples to tackle the problem of frequently misrecognized words.

3. OUTLINE OF THE ALGORITHM

We modified our pre-trained JANUS[1] speech recognizer for the given task to run as a phoneme recognizer with smoothed phoneme-bigrams. We will need both the phoneme and the speech recognizer to perform our algorithm.

We will not need any fine-labeled speech data, but we will need transcriptions on a word-level, as they are needed for training a speech recognizer. Additionally we will need the following prerequisites:

Prerequisites:

1. create word labels for the whole training set by running the existing speech recognizer on all training utterances, resulting in the word boundaries for all word occurrences.
2. create a phoneme confusability matrix for the underlying speech recognizer
3. create a smoothed phoneme language model
4. analyze frequent misrecognitions of the underlying SR engine on training and cross validation set.
5. from this generate a list of word tuples which should be modeled in the dictionary

Analyzing the misrecognitions of our speech recognizer, we found that they were often due to misrecognition of short words. The term "short words" includes words which have "short" pronunciations. Another problem was caused by words which became confusable after looking at the possible pronunciation variants (e.g. the German words "ist","es" in Table 5). Introducing word tuples for modeling such words within their context increases speech recognition performance, as it reduces both acoustic and language model confusability.

Using both, the speech and the phoneme recognizer, Dictionary Learning can be performed by the following
Dictionary Learning Algorithm:

1. collect all occurrences of each word/tuple in the database and run the phoneme recognizer on them using the smoothed phoneme LM
2. compute statistics of the resulting phonetic transcriptions of all words/tuples
3. sort the resulting pronunciation candidates using a confidence measure and define a threshold for rejecting statistically irrelevant variants
4. reject variants that are homophones to already existing dictionary entries
5. reject variants which only differ in confusable phonemes
6. add the resulting variants to the dictionary
7. test with the modified dictionary on the cross validation set (optional)
8. retrain the speech recognizer, allowing the use of multiple pronunciations during training.
9. as an optional step corrective phoneme training can be performed
10. test with the resulting recognizer and the modified dictionary on the cross validation set
11. create a new smoothed language model for the phoneme recognizer, incorporating all new variants.
12. optional second pass

In step 5 the phoneme confusability matrix is used to reject variants which differ only in phonemes which are confusable to the recognizer and therefore would lead to erroneous training of confusable phonemes (eg. reject variant D A M vor the German word "dann" if the phonemes N and M are highly confusable). This avoids further contamination of the underlying phonetic units. Step 8 leads to more accurate training data and to a better discrimination of the phonetic units. The new phoneme language model, computed in step 11, incorporates statistical knowledge (similar to phonetic rules) about already observed phoneme sequences, and should be used the next time this algorithm is applied.

4. EXPERIMENTAL SETUP

4.1. Database and Baseline System

All experiments within this paper were performed on a German database called the German Spontaneously Scheduling Task (GSST), which is collected as a part of the VERBMOBIL project. In this task human-to-human spontaneous dialogs are collected at four different sites within Germany. Two individuals are given different calendars with various appointments already scheduled and have to find a time slot which suits both of them. The test vocabulary contained more than 3300 entries.

	Training	Test
#Dialogues	608	8
#Utterances	10735	110
#Words	281160	2346
Vocabulary Size	5442	543

Table 1: GSST Database

For the experiments reported here we used the hybrid LVQ/HMM recognizer of JANUS 2, our spontaneous speech-to-speech translation system [2, 3], using 69 context independent[1] phoneme models, including noise models.

4.2. Experiments

In our first set of experiments we carried out all the steps described in the previous section, with exception of retraining. Table 2 summarizes the first results and their comparison with the baseline system that does not use alternative pronunciations. In experiment A1 we generated alternative pronunciations which do not result in homophones in the dictionary. In experiment A2 we additionally used the phoneme confusability matrix to reject variants which differ only in phonemes which were confusable to the recognizer.

For the second set of experiments we used a slightly improved baseline system. Table 3 summarizes the results after re-training and the comparison with the baseline system B that does not use alternative pronunciations. In experiment B1 we generated alternative pronunciations as in experiment A2. In experiment B2 we additionally used discriminative phoneme training to increase the discrimination between confusable phonemes.

[1] Our currently best spontaneous speech recognizer on GSST/VERBMOBIL (PP 62, approx. 3600 word dictionary) performs at a word accuracy of about 74.6% on the official 1995 VERBMOBIL evaluation set.

dictionary used	WA	error reduction
baseline system A[a]	60.8%	—
experiment A1[b]	63.5%	4.4%
experiment A2[c]	64.2%	5.6%

[a] no alternative pronunciations were used
[b] alternative pronunciations, but no homophones
[c] variants with confusing phonemes were rejected

Table 2: Recognition results using Dictionary Learning

dictionary used	WA	error reduction
baseline system B[a]	61.7%	—
experiment B1[b]	64.9%	5.2%
experiment B2[c]	65.6%	6.3%

[a] no alternative pronunciations were used
[b] same as A2, retraining without step 9
[c] same as A2, retraining with step 9

Table 3: Recognition results after re-training

Retraining the speech recognizer with the new dictionary improved the overall recognition performance; additional discriminative phoneme training gave further improvements in recognition performance.

In a third set of experiments (C1,C2,C3) we examined the most frequent words/tuples and used the Dictionary Learning algorithm to generate pronunciations for them. No retraining was performed in this experiment, so further improvements after re-training are likely. The increased recognition performance of the baseline system is due to the use of trigram language models in these experiments. The dictionary of the baseline system C had 3309 entries. In experiment C1 additional 119 tuples were added to the dictionary. System C2 used 130 variants of words and system C3 used 297 variants for words and tuples.

dictionary used	WA	error reduction
baseline system C[a]	65.4%	—
experiment C1[b]	67.5%	3.1%
experiment C2[c]	67.7%	3.4%
experiment C3[d]	68.4%	4.4%

[a] no alternative pronunciations were used
[b] using 122 word tuples, no variants
[c] no tuples, but variants
[d] using 122 word tuples and variants

Table 4: Recognition results with word tuples (no retraining)

The experiments with word tuples have shown that the pronunciation variants found model dialectic variations as well as coarticulation of short words in a larger word context.

4.3. Examples

Some examples for resulting pronunciations for word tuples are shown in the following two tables. In the first table you see pronunciation variants for the German words "ist" and "es" and for the contraction of the two words, resulting in the tuple "ist_es". The second table shows pronunciation candidates for the tuples "einen_Termin" and "noch_einen_Termin", two tuples which occur very often in the given task and which are pronounced very sloppy – resulting in quite a lot pronunciation variants which represent dialectic variations which can often be found in spontaneously spoken German speech.

occurrences	pronunciations
23.35 %	**? I S T**
36.55 %	? I S

Pronunciation Candidates for "ist"

occurrences	pronunciations
11.40 %	S
21.24 %	**? E S**
23.83 %	? I S

Pronunciation Candidates for "es"

rank	pronunciations
(1)	? I S I S
(2)	? I S E S

Pronunciation Candidates for "ist es"

Table 5: Example 1

rank	pronunciations
(1)	? AI N T ER M IE N
(2)	? AI N E2 N T ER M IE N
(3)	N T ER M IE N
(4)	N E2 N T ER M IE N
(5)	? AI N E2 N T ER M IE N
(6)	? E N T ER M IE N

Pronunciation Candidates for "einen Termin"

rank	pronunciations
(1)	N O X AI N T ER M IE N
(2)	N O X ? AI N T ER M IE N
(3)	N O X AI N E2 N T ER M IE N
(4)	N O X E2 N T ER M IE N

Pronunciation Candidates for "noch einen Termin"

Table 6: Example 2

5. CONCLUSIONS

We have pointed out that adding or modifying phonetic variants by hand is an error prone and labor intensive procedure. We gave the outline of a data-driven algorithm for Dictionary Learning which enables us to automatically generate new entries to a phonetic dictionary in a way that all entries are consistent with the underlying phonetic modeling. We showed that some of the frequently misrecognized words can be modeled more accurately by using word tuples and that pronunciations for such tuples can also be found using Dictionary Learning. Using smoothed phoneme language models during the phoneme recognition enables us to incorporate statistical knowledge about previously observed phoneme sequences without having to keep track of and to apply phonological rules. Our experiments showed that our Dictionary Learning algorithm for adapting and adding phonetic transcriptions to existing dictionaries improves the overall recognition performance of the speech recognizer significantly.

ACKNOWLEDGEMENTS

This research was partly funded by grant 413-4001-01IV101S3 from the German Ministry of Science and Technologie (BMBF) as a part of the VERBMOBIL project. The views and conclusions contained in this document are those of the authors. The author wishes to thank all members of the Interactive Systems Laboratories for all the useful discussions and active support, especially Michael Finke and Monika Woszczyna for their helpful discussions, and Klaus Ries for assistance with the word tuple language models. Special thanks to my advisor Alex Waibel.

6. REFERENCES

1. Tilo Sloboda: *Dictionary Learning: Performance through Consistency*, Proceedings of the ICASSP 1995, Detroit, volume 1, pp 453-456.
2. A.Waibel, M.Finke, D.Gates, M.Gavaldà, T.Kemp, A.Lavie, L.Levin, M.Maier, L.Mayfield, A.McNair, I.Rogina, K.Shima, T.Sloboda, M.Woszczyna, T.Zeppenfeld, P.Zhan: *JANUS II — Translation of Spontaneous Conversational Speech*, Proceedings of the ICASSP 1996, Atlanta, volume 1, pp 409-412.
3. M. Woszczyna, N. Aoki-Waibel, F.D. Buø, N. Coccaro, K. Horiguchi, T. Kemp, A. Lavie, A. McNair, T. Polzin, I. Rogina, C.P. Rose, T. Schultz, B. Suhm, M. Tomita, A. Waibel: *JANUS 93: Towards Spontaneous Speech Translation*, Proceedings of the ICASSP 1994, Adelaide, volume 1, pp 345-348.
4. M.Woszczyna, N.Coccaro, A.Eisele, A.Lavie, A.McNair, T.Polzin, I.Rogina, C.P.Rose, T.Sloboda, M.Tomita, J.Tsutsumi, N.Aoki-Waibel, A.Waibel, W.Ward: *Recent Advances in JANUS, a Speech to Speech Translation System*, Proceedings of the EUROSPEECH, Berlin, 1993.
5. J.L.Gauvain, L.F.Lamel, G.Adda, M.Adda-Decker: *The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA Wall Street Journal Task*, Proceedings of the ICASSP 1994, Adelaide, volume 1, pp 557-560.
6. Toru Imai, Akio Ando, Eiichi Miyasaka: *A New Method for Automatic Generation of Speaker-Dependent Phonological Rules*, Proceedings of the ICASSP 1995, Detroit, volume 1, pp 864-867.