DICTIONARY LEARNING: PERFORMANCE THROUGH CONSISTENCY

Tilo Sloboda
sloboda@ira.uka.de

Interactive Systems Laboratories
University of Karlsruhe — Germany
Carnegie Mellon University — USA

ABSTRACT

We present first results from our efforts in automatically increasing and adapting phonetic dictionaries for spontaneous speech recognition. Spontaneous speech adds a variety of phenomena to a speech recognition task: false starts [1], human and nonhuman noises [2], new words [3] and alternative pronunciations. All of these phenomena have to be tackled when adapting a speech recognition system for spontaneous speech. For phonetic dictionaries (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 it might be found in a lexicon. Additionally modifications of the dictionary should not lead to a higher phoneme confusability. Therefore we propose a data-driven approach to add new pronunciations to a given phonetic dictionary, in a way that they model the given occurrences of words in the database. We show how even a simple approach can lead to significant improvements in recognition performance. First experiments have been performed on the German Spontaneous Scheduling Task (GSST), using the speech recognition engine of JANUS-2 [4, 5, 6], the spontaneous speech-to-speech translation system of the Interactive Systems Laboratories at Carnegie Mellon and Karlsruhe University.

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 doesn't have to be the most frequent variant for a given task (e.g. in spontaneous speech), nor does it necessarily yield the best recognition performance, given the current acoustic modeling. If the phonetic transcriptions in the dictionary don't match the actual occurrences in the database, the phonetic units will be contaminated during the training with inadequate acoustics. This will degrade the overall performance of the recognizer.

State-of-the-art speech recognition systems (e.g. [8, 7]) start to put more and more effort into creating adequate dictionaries with alternative pronunciations and function words, which can also model interword effects such as coarticulation between words. This is usually done either by modifying the dictionary by hand or applying phonological rules 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 doesn't aim at increasing the overall system performance. Furthermore it is error prone – all kind of 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 (as it could be found in a lexicon) – this isn't necessarily the most frequent or even the most likely transcription for a given task.
- alternative 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.
- as it is hard to say which variants are statistically relevant for a given task, the maintainer of the dictionary can easily miss a relevant form.

Therefore we propose a data-driven approach to improve existing dictionaries and automatically add new 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.

In our experiments we showed that even using a simple algorithm to extract candidates for phonetic variants yields a significant increase in recognition performance. More sophisticated algorithms yield even better performances.

2. DICTIONARY LEARNING

We will give an outline of two algorithms for Dictionary Learning. The first algorithm aims at improving the recognition performance of a given speech recognizer without retraining; the second algorithm is aiming at optimizing the dictionary for retraining, so that contaminated phonetic units will get more accurate training.

Dictionary Learning can also be used to add new words to the dictionary – this is clearly less work than adding them by hand. Getting a good initial estimate for the pronunciations of infrequent words is a separate problem. This can be achieved by online input of extra samples for these words. This extra input can then be used for estimating the pronunciation (but not for training the acoustic models). Once there are enough samples for a word in the database, the pronunciations should be build on those samples only.

Applying Dictionary Learning whenever a larger amount of new data is added to the database will also help to keep the dictionary consistent.

2.1. Outline of Algorithm A

We modified a pre-trained speech recognizer for the given task to run as a phoneme recognizer with smoothed phoneme-bigrams (e.g. based on our JANUS speech recognition engine [4, 5, 6] in context independent mode[1]). Using this setup, Dictionary Learning can be performed by the following algorithm:

1. create word labels for the whole training set and a phoneme confusion matrix for the underlying speech recognizer
2. collect all appearances of each word in the database, run the phoneme recognizer on them and compute a statistic of the resulting phonetic transcriptions of each word
3. define a cutoff point for rejecting statistically irrelevant variants
4. avoid further contamination of the underlying phonetic units, by using the phoneme confusion matrix of the speech recognizer to reject variants which would lead to erroneous training of confusable phonemes (e.g. reject variant D A M vor the German word "dann" if the phonemes N and M are highly confusable)
5. test with the modified dictionary

2.2. Results of Algorithm A

For the experiments reported here we used the hybrid LVQ/HMM recognizer of JANUS [6], using 69 context independent[1] phoneme models (including noise models [2]) as a baseline system. We used a subtask of the German Spontaneous Scheduling Task (GSST), with a training set of 1967 distinct words and a test set of 496 distinct words.

In the experiments A1 and A2 we carried out all the steps described in the previous section.

The following 4 examples show alternative pronunciations which were found by the algortihm. The pronunciations which are printed in bold face are the ones which were allready in the dictionary.

occurrences	pronunciations
2.86 %	Z EH N E2 N AHR
8.57 %	TS EH M I N AHR
8.57 %	Z E M IE N AHR
8.57 %	**Z E M I N AHR**
11.43 %	TS EH M IE N AHR
14.29 %	Z EH M I N AHR
14.29 %	Z EH M IE N AHR

Pronunciation Candidates for "Seminar"

occurrences	pronunciations
2.84 %	D EH N S T AH
2.84 %	D IE N S T AH
2.84 %	D IE N SCH T AH K
4.96 %	D IE N S T AH X
12.77 %	D EH N S T AH K
38.30 %	**D IE N S T AH K**

Pronunciation Candidates for "Dienstag"

occurrences	pronunciations
3.96 %	K OE N E2
5.94 %	**K OE N E2 N**
72.28 %	K OE N

Pronunciation Candidates for "können"

occurrences	pronunciations
8.49 %	? A N
16.04 %	**? AI N E2 N**
55.66 %	? AI N

Pronunciation Candidates for "einen"

Indeed if one follows the actual paths in the two figures, one finds dialectic variations of the given German words.

dictionary used	WA	error reduction
baseline system[a]	60.8 %	—
experiment A1[b]	63.5 %	6.9%
experiment A2[c]	64.2 %	8.7%

[a] no alternative pronunciations were used

[b] alternative pronunciations, but no homophones

[c] variants with confusing phonemes were rejected

Table 1. Recognition results using Dictionary Learning

Table 1 summarizes the results and their comparison with the baseline system that doesn't use alternative pronunci-

[1] Our currently best spontaneous speech recognizer on GSST (PP 70, approx. 2000 word dictionary) uses context dependent phoneme models and performs at a word accuracy of about 70%.

ations. In the first experiment (A1) we generated alternative pronunciations which don't result in homophones in the dictionary. In the second experiment (A2) we additionally used the phoneme confusion matrix to reject variants which differ only in phonemes which are confusable to the recognizer.

Adding alternative pronunciations which were generated by Dictionary Learning gave a significant improvement in performance.

occurrences	pronunciations
1.65 %	D E2 N I CH
8.68 %	**N I CH T**
9.50 %	N I CH
12.81 %	? I CH T
14.88 %	M I CH T
15.70 %	M I CH
16.12 %	? I CH

Table 2. Pronunciation Candidates for ”nicht”

Table 2 shows an example of variants which differ only in highly confusable phonemes. Inconsistencies in the original dictionary can lead to such confusion pairs.

In the next section we will show how retraining the recognizer with the new dictionary improves the overall performance and the discrimination between confusable phonemes.

2.3. Outline of Algorithm B

For retraining the recognizer using the new dictionary with alternate pronunciations, the following steps have to be performed additionally:

1–5 same as in algorithm A

6. retrain the spontaneous speech recognizer, allowing the use of multiple pronunciations during training. This leads to more accurate training data for the phonetic units and to a better discrimination of the phonetic units
7. optional step: corrective training of pronunciations of a word which only differ in highly confusable phonemes (e.g. variants M I CH T and N I CH T of the German word ”nicht” are trained discriminatively, as they only differ in the highly confusable phonemes N and M).
8. test with the resulting recognizer and the modified dictionary

Step 7 aims at additionally performing discriminative phoneme training between pairs of highly confusable phonemes and can be performed optionally.

2.4. Results of Algorithm B

For the second set of experiments we used a slightly improved baseline system, which used another LDA transformation.

Table 3 summarizes the results after re-training and the comparison with the baseline system that doesn't use alternative pronunciations. In the first experiment (B1) we generated alternative pronunciations as in experiment A2. In the second experiment (B2) we additionally used discriminative phoneme training to increase the discrimination between confusable phonemes.

dictionary used	WA	error reduction
baseline system[a]	61.7 %	—
experiment B1[b]	64.9 %	5.1%
experiment B2[c]	65.6 %	6.3%

[a] no alternative pronunciations were used

[b] same as A2, retraining without step 7

[c] same as A2, retraining with step 7

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.

3. CURRENT WORK

We are currently working on evaluating our algorithm on other tasks, such as Wall Street Journal (WSJ) and English Spontaneous Scheduling Task (ESST).

4. 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. By our experiments we have shown that our algorithm for adapting and adding phonetic transcriptions to a dictionary improves the overall recognition performance of a speech recognizer.

5. ACKNOWLEDGEMENTS

This research was partly funded by grant 413-4001-01IV101S3 from the German Ministry of Science and Technology (BMFT) 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 useful discussions and active support. Special thanks to my advisor Alex Waibel, to Ivica Rogina and Monika Woszczyna.

REFERENCES

[1] Douglas O'Shaughnessy: *Correcting Complex False Starts in Spontaneous Speech*, Proceedings of the ICASSP 1994, Adelaide, volume 1, pp 349-352.

[2] Tanja Schultz, Ivica Rogina: *Acoustic and Language Modeling of Human and Nonhuman Noises for Human-to-Human Spontaneous Speech Recognition*, Proceedings of the ICASSP 1995.

[3] B.Suhm, M.Woszczyna, A.Waibel: *Detection and Transcription of New Words*, Proceedings of the EUROSPEECH, Berlin, 1993.

[4] L. Osterholtz, A. McNair, I. Rogina, H. Saito, T. Sloboda, J. Tebelskis, A. Waibel, and M. Woszczyna: *Testing Generality in JANUS: A Multi-Lingual Speech to Speech Translation System*, Proceedings of the ICASSP 1992, San Francisco, volume 1, pp 209–212.

[5] 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.

[6] 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.

[7] 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.

[8] Chuck Wooters, Andreas Stolcke: *Multiple–Pronunciation Lexical Modeling in a Speaker Independent Speech Understanding System*, Proceedings of the ICSLP 1994, Yokohama, volume 3, pp 1363-1366.