Improving Automatic Speech Recognition Accuracy by Means of Pronunciation Modeling

Vladimir Chuchupal, Anton Korenchikov

Mathematical Problems of Pattern Recognition Departament,
Dorodnicyn Computing Centre, Russian academy of sciences

chuchu@ccas.ru

Abstract

We explore the properties of the pronunciation variation (PV) models as an approach for automatic speech recognition accuracy improvement. The PV model is formally defined as well as the methods of its parameter estimation. We show that utilizing PV models could substantially increase the accuracy of automatic recognition of natural speech.

1. Introduction

The pronunciation of a word in a speech recognition system (ASR) is usually determined by its pronunciation or phoneme transcription. As a rule most of the words has a single pronunciation transcription, namely the basic or canonical one.

In spontaneous speech pronunciation may substantially differ from the canonical one and this is one of the most important sources of errors of the speech recognizer.

There are currently two approaches to pronunciation variation (PV) modeling for ASR [1, 2]. Explicit modeling describes all probable pronunciation variations in terms of explicit changes of the basic word transcriptions. In other words, in explicit modeling the given word pronunciation could be defined as a set of most probable word transcriptions. Implicit modeling [3] describes variations in pronunciation by means of changes in phoneme models.

The both approaches do not eliminate the need to use the basic transcriptions.

The correct implementation of PV models may have a great impact on the accuracy of ASR. Such a conclusion follows from the heuristic analysis of the errors made by the ASR as well as the oracle-style experiments. As it was shown in [4] the use of adequate phonemic transcriptions can reduce word error rate (WER) approximately one half.

The reported improvements in WER obtained with PV models in experiments on known data corpora are still far from the expected ones. In [1] on Dutch corpora VIOS the WER decreased at 0.8% from 10.7% to 9.9%, with 4.9 pronunciations per a vocabulary word. In [3] on Switchboard corpora the implementation of implicit PV models led to WER improvement of 1.7% (from 39.4% to 37.7%). In [5] on NIST–2000 Hub-5 data the use of pronunciation variation models improved WER by 2.2%: from 54.6% to 52.4%.

In this study we implemented the pronunciation variation model in the existing Russian ASR. We follow the explicit approach to pronunciation variation modeling in that all changes in pronunciation could be adequately described in terms of deletions, substitutions and insertions of phonemes.

For implementation of this approach we have to address the following issues:

- define PV models,
- find most probable phone transcriptions for words,
- estimate the parameters of PV models,
- embed PV models in the search procedure.

2. Pronunciation variation model

Let $X = \{x_t\}, t = 1, \ldots, T$ be a sequence of the vector parameters of the observed speech signal, $W = \{w_i\}, i = 1, \ldots, N$ be a sequence of the vocabulary words. Then the most probable word sequence W^* can be obtained from the equation [6]

$$W^* = \arg\max_W \frac{P(X|W)P(W)}{P(X)}. \qquad (1)$$

The first factor $P(X|W)$ in the numerator (1) is a data likelihood. It could be obtained with the help of the acoustic phone models. The value of the second factor $P(W)$ is estimated with the help of the language model.

We use t^w to denote the phonemic transcription (pronunciation model) of a word w. The set of phonemic transcriptions of the given word w is denoted as T^w. The pronunciation model for the word sequence W is denoted T^W. The designation t^W will be used as a notation for an element of T^W.

The conventional speech decoding and recognition procedures as a rule define the best sequence of acoustical

models (phonemic transcriptions), not the best sequence of word. That is used de facto used instead of (1):

$$t^{W^*} = \underset{t^W}{arg\,max}\, \frac{P(X|t^W)P(t^W)}{P(X)}. \quad (2)$$

Then the most probable word sequence could be obtained by mapping each pronunciation model into the corresponding word, i.e.:

$$t^{W^*} \rightarrow W^*. \quad (3)$$

If for all words there is a single pronunciation per word in vocabulary the methods (1) and (2) are equivalent.

Using the equality $P(t^W) = P(t^W|W)P(W)$ the expression (2) could be written as:

$$W^* = \underset{t^W}{arg\,max}\, \frac{P(X|t^W)P(t^W|W)P(W)}{P(X)}. \quad (4)$$

The expression (4) differs from that in (2) in that it contains the factor $P(t^W|W)$ that accounts the pronunciation variation. The set of probabilities $P(T^W|W) = \{P(t^W|W), t^W \in T^W\}$ is considered as parameters of the PV model.

3. Estimation of parameters of pronunciation variation model

In order to use (4) we need to know the parameters of three models: acoustic, language and pronunciation ones.

The language model parameters for estimation $P(W)$ usually considered as independent of the acoustic models. Therefore the estimation of language model parameters could be performed in the independent manner exactly as it is done in conventional (8) approach.

The pronunciation model parameters $P(T^W|W)$ are dependent on the acoustic training data, therefore the independent (of acoustic one) estimation of $P(T^W|W)$ is not correct.

Consider the maximum likelihood estimate of the pronunciation model parameters.

Suppose that the training corpora X is such that for all its utterances we know the sequence of words $w_1 w_2 \dots w_N$ as well as a sequence of the phonemic transcriptions $t_1^w t_2^w \dots t_N^w$. In such a case the most probable estimate of the parameters $p(t^w|w)$ will be obtained by solving the following:

$$p(t^w|w) = \underset{w,t^w}{arg\,max} \prod_{w,t^w} p(t^w|w). \quad (5)$$

This frequency estimate is similar to the estimate for the n–gram language model[8]:

$$p(t^w|w) = \frac{\#\{t^w\}}{\#\{w\}}, \quad (6)$$

where $\#$ denotes the number of events in curly braces, encountered in the training data. Therefore the most probable estimate for the given transcription will be the relative frequency of that transcription in the training corpora.

Since the independent estimation of the acoustic and pronunciation model parameters is not correct consider the algorithm consisting of two-step iterations.

Suppose that there are training speech corpora along with the vocabulary and for each vocabulary word we know all of the pronunciation variants. Consider for a start the variants are equally probable.

In the first step the maximum likelihood estimates of the acoustic model parameters are obtained. The conventional training methods based on forward-backward and Baum-Welch algorithms can be used.

In the second step using (6) the maximum likelihood estimations of PV model parameters are obtained. It is done with the help of the co-called "restricted" recognition of all utterances in the speech corpora. The term "restricted" means that the true word sequence is known in advance and the target is to find out the most probable sequence of phonemes or, in other words, the most probable sequence of transcriptions.

These steps are repeated either for the fixed number of times or until a stopping criteria will be met.

4. Embedding of the pronunciation variation model into a speech decoder

A conventional way to use several pronunciation transcriptions per word in a speech decoder consists of inclusion of each transcription into the pronunciation vocabulary and handling this transcription in an independent manner as if it is a transcription of a new word. This approach implies no changes in the search algorithms (2)-(3).

It is not the optimal solution though.

Rewrite the expression (1):

$$P(W|X) = \frac{\sum_{t^W \in T^W} P(X|t^W)P(t^W)}{P(X)}. \quad (7)$$

From (4) and (7) it follows that the most probable sequence of words W^* should satisfy

$$W^* = \underset{W}{arg\,max} \sum_{t^W \in T^W} P(t^W|X)P(t^W). \quad (8)$$

Equality (8) allows us to define the most probable word sequence (not the most probable phoneme or transcription sequence) that is exactly what we need of the speech recognition system.

Decision (8) differs from the one of (2)-(3) in that we need to take into account the relative frequencies of word phoneme transcriptions and make the final decision using the weighted sum of the transcription likelihoods.

The practical implementation of (8) is associated with the drawback because of lexicon tree pruning [8]. Some leaves of the tree have been pruned because of the relatively small likelihoods. In such a case the likelihoods of these leaves are not known and the corresponding transcriptions could not be used.

To overcome that drawback, consider the following version of (8):

$$W^* = \underset{W,t^W}{arg\,max}\, P(t^W|X)P(t^W). \qquad (9)$$

Here the weighted sum of the likelihoods is replaced with the likelihood of the most probable transcription weighted by $P(t^W|X)$.

5. Numerical experiments

The performance of the considered PV models have been compared on the speech corpora ISABASE–2 [9] and TeCoRus [10]. The training data of the first test consisted of speech utterances of 200 speakers of ISABASE–2 (40K utterances) and 50 speakers from TeCoRus (3K utterances). The test material consisted of the 776 utterances that contained connected digit strings (3147 digits). The vocabulary has been limited to the digits. The reason to use numbers was that the numbers and numerals could provide a lot of examples of pronunciation variations.

No language models have been used.

The recognition results in terms of word error rate (WER) values are presented in Table 1. The column "Basic" contains the results for the case when the basic transcription is used only. The column "Conv." corresponded to the method (1-3). The column "Opt." contains the results for the method (8). The column "SubOpt." contains results for the method (9). The row "Variability" contains the mean number of transcriptions per vocabulary word.

Table 1: *Word Error Rate for some pronunciation variation models (TeCoRus data only).*

Method	Basic	Conv.	Opt.	SubOpt.
WER	1.62	5.78	2.00	3.17
Variativity	1.0	1.9	1.9	1.9

The results depicted above could be interpreted as an evidence of lack of pronunciation variability in the test corpora. It can be true because the speakers of TeCoRus belong to the same high-educated professional group and were living in Moscow region. The test material contained read and carefully articulated speech.

The lack of the PV in the first test could explain the observed behavior of the training algorithm: on the TeCoRus data with the increasing number of iterations the mean number of transcriptions per word come down to one.

To obtain recognition results for the data with actual pronunciation variability the second recognition experiment has been performed. The training set of the second test was the same as in the first test. The test set consisted of 867 utterances of 11 test speakers of TeCoRus. These data mostly consists of the sequences of digits and numerals. The vocabulary of the test set consisted of 129 words. Test utterances also contained additive noises as well as disfluencies that typically led to the recognition errors.

The pronunciation vocabulary contains 129 numerals.

Table 2 shows the WER values for the second test. The table column "Conv." shows the WER value for the case when the basic pronunciations were used only.

Table 2: *WER value for TeCoRus extended data.*

Method	Basic	Conv.	Opt.	SubOpt.
WER	7.78	7.57	7.38	7.44
Variativity	1.0	1.3	1.3	1.3

The results drawn in (2) could be considered as more relevant to the expected ones. The best approach appears to be the one that corresponds to the frequency weighting of the pronunciation variants(8). The approach with the inclusion the alternative transcription into the pronunciation vocabulary (1 - 3) appears to be less effective for the both the (8) and (9) algorithms. In all cases the inclusion of pronunciation variations appears to be more effective than the use of basic transcriptions only.

The WER improvements in the second test were not as substantial as it could be expected though. On the one hand it could be because of the type of the test material. At the same time the WER improvements observed might be due to the fact that the speech corpora TeCoRus and Plantronics had been collected in different conditions. TeCoRus had been recorded with a Senheiser professional microphone while ISABASE–2 corpora had been recorded with a Plantronics microphone.

To clarify these issues the third recognition test has been performed on the speech corpora that contained natural spontaneous speech extracted from radio interviews. We used the interviews downloaded from the radio station "Echo Moscow" [11].

The initial set of pronunciation transcriptions for numerals as well as their relative frequencies were the same as in the previous test.

The interviews were automatically segmented. Then the utterances with the numerals were found and extracted as separate speech files. The test set consisted of 200 speech utterances of 2–4 words each, with total vocabulary of 91 words.

No language models were used during recognition.

Table 3 presents the results for this test. The table column "Equal." contains the WER values for the method(8)

in the case when the equal relative frequencies for all competitive transcriptions were used.

Table 3: *WER values for interview data.*

Method	Basic	Conv.	Opt.	SubOpt.	Equal
WER	69.3	57.44	59.7	60.0	59.5

The substantially higher WER values obtained because of the lack of the language model, mismatch between training and testing conditions for acoustic models, and noisy environment during interviews.

In the third test the observed relative improvements in WER was from 13,4% to more than 17,1% comparing to 5% relative improvement in the previous test.

It is shown therefore that for fluently spoken numerals the use of PV models can lead the substantial improvement of the speech recognition accuracy.

Note that there is another (besides of pronunciation changes) possible reason of improvements of the accuracy in the third test. There is a significant mismatch in the training and testing data for the test data were coded in MP3 format. However if it was the case then the similar WER improvements were to take place in the second test. It had not happened though.

The observed absence of improvement in WER (compared with the other methods) for the methods with weighting of competitive transcriptions can be explained with regard to the language modeling. The transcription weighting as well as using the number of competitive word transcriptions for numerals has an effect that is similar using the unigram language model. In the test material the relative numeral frequencies were much higher than in the other. The use of conventional method has an effect of using bigger unigram weights for numerals that were relevant to the data of the test corpora.

6. Conclusion

The research of the methods for improving the automatic speech recognition accuracy through the use of pronunciation variation models is fulfilled. The probabilistic pronunciation variation model is formulated and well as the ways to estimate the model parameters. The numerical experiments shows that the implementation of the pronunciation variation models is an effective way to improve accuracy of spontaneous speech recognition.

7. Acknowledgments

The work was supported by Russian Fund of Basic Research, project 14-01-00607

8. References

[1] Wester M. Pronunciation modeling for ASR knowledge-based and data-derived methods // Computer Speech and Language. 2003. Vol. 17, P. 69-85.

[2] Fosler-Lussier E. Dynamic pronunciation models for automatic speech recognition. Ph.D. thesis. University of California, Berkley, CA, 1999.

[3] Saraclar M., Khudanpur S. Pronunciation change in conversational speech and its implications for automatic speech recognition // Computer Speech and Language. 2004. Vol. 18(4). P. 375-395.

[4] Saraclar M., Nock H., Khudanpur S. Pronunciation modeling by sharing Gaussian densities across phonetic models // Computer Speech and Language. 2000. Vol. 14(4). P. 137-160.

[5] Zheng J., Franco H., Stolcke A. Modeling word-level rate-of-speech variation in large vocabulary conversational speech recognition // Speech Communication. 2003. Vol. 41. P. 273285.

[6] Jelinek F. Statistical Methods for Speech Recognition. Cambridge, Massachusetts: The MIT Press, 1997.

[7] Chow Y.-L., Schwartz R. The N-Best Algorithm: Efficient Procedure for Finding Top N Sentence Hypotheses // Proceedings of the International Conference on Acoustic, Speech and Signal Processing, ICASSP. 1990, P. 199-202.

[8] Young S., Bloothooft G. (Eds.). Corpus-based methods in language and speech processing. Dordrecht: Kluwer Academic Publishers (Text, Speech and Language Technology series, Vol. 2), 1997.

[9] Bogdanov D. S., Krivnova O. F., Podrabinovitch A. J., Arlazarov V;L. Creation of Russian Speech Databases: Design, Processing, Development Tools // Proceedings of the International Conference on Speech and Computers, SPECOM. Moscow, 2004.

[10] Chuchupal V.J., Makovkin K.A., Chichagov A.V., Kuszetsov V.B., Ogarysyev V.F. Speech corpora TeCoRus. Data base registration certificate 2005620205, 2005.

[11] http://www.echo.msk.ru.